# Chapter 1

# Introduction

This dissertation reports the recognition of off-line handwritten characters of Meitei Mayek script. It is a newly revived script which otherwise had remained dormant for more than three centuries. Very few of the languages spoken in the north eastern part of India has its own script, Meitei Mayek is one of them. It is the script for Manipuri language (Meiteilon), the language predominantly used in the state of Manipur.

## 1.1 Handwritten Character Recognition (HCR)

Optical Character Recognition (OCR) is the technique by which computers convert printed or handwritten texts into a form which can be used within its processing applications [212]. Handwritten Character Recognition (HCR) is a sub-field of OCR where the texts concerned are handwritten ones. HCR can be off-line where handwritten texts in an image are converted into letter codes which are usable within computer and text-processing applications; or on-line which is the automatic conversion of text as it is written on a special digitizer or PDA. The research work presented in this thesis focuses on off-line HCR and therefore, it is worth mentioning that by HCR, it would mean off-line HCR for the rest of the thesis. HCR has been studied for a long time by researchers all over the world. The problem of HCR is challenging because of certain particular characteristics of different scripts and because of the variations in the writing styles of different individuals. Also, there are variations in the handwriting of a single person due to the mental state they are in while writing. For example, a person while in a calm state of mind will have a smooth handwriting, whereas the same person in a tensed state has the possibility to write in a different way resulting in a different handwriting. So therefore, the degree of variations in handwritten texts is enormous.

Organizations such as those in banking sectors, health care industries, and many others deal with handwritten documents extensively daily basis in the form of bank checks, medical forms, etc. It is a common practice to digitize handwritten documents so that they can be stored more compactly, and can be electronically edited and searched. However, the amount of such documents present today is so huge that it becomes a tedious and time-consuming job to manually process them. Hence, the development of an automatic-reading system is the key solution even if their recognition accuracy is not as good as humans. As pointed out in [153], the text data stored in printed or handwritten documents are of immense importance for the purpose of future reference and serves as a record for history, culture, literature, etc.

Studies in HCR have focused mainly on major scripts for languages like English, Chinese, Arabic and Japanese. Several studies have also been reported in off-line HCR of some Indian scripts; Bangla [32, 151, 163, 174], Devanagari [17, 145, 197], Gujarati [50, 51, 154], Kannada [52, 142, 168, 197], Gurmukhi [4, 103, 198], Oriya [33, 152, 173, 218], Telugu [44, 160, 189, 202], Malayalam [124, 150], Tamil [97, 164, 193] and Nastaliq (Urdu) [6, 12, 81]. Maximum research on off-line handwritten recognition has been done for Bangla script.

### 1.1.1 Applications of HCR

The applications of HCR are widespread. It plays an important role in the following application fields:

- Text-to-Speech conversion: Handwritten text is converted to machine codes which the computers understand and process. The output is then fed to a speech synthesizer for conversion to speech. This also acts as a reading aid for the blind.
- Machine translation: Handwritten documents such as poems, articles, stories, etc. written in one language can be converted into another target language.
- Machine transliteration: For scripts like Meitei Mayek which was not in used for a long period of time until it got revived, the handwritten documents

written in Bangla script (which was in use before Meitei Mayek got revived) can be converted to Meitei Mayek for the younger generation who do not know Bangla script.

- Automatic reading of postal addresses, bank check amounts, and forms which tremendously reduces manual interpretation.
- Automatic entry of textual information on images into computers by digitization which further enables automatic searching and editing.
- Content-based image retrieval: a word can be queried in and the system retrieves the digitized hand documents which contain the word in the query.
- Automatic answer sheets evaluation.
- Keyword spotting: the system retrieves a word with the highest frequency of occurrence in a handwritten document image and thus provides a keyword of the document image.

### 1.1.2 Challenges in development of HCR system

- Unavailability of datasets: For certain scripts which are new such as Meitei Mayek, the unavailability of datasets is the first challenge faced by researchers. In order to carry out HCR on a script, creation of datasets is a prerequisite. This becomes a challenge for Meitei Mayek as the script is a low-resourced one and not many documents are available which can be used for the dataset creation.
- Varying handwritings: There exists great variations in the handwritings of different individuals. This introduces high intraclass dissimilarity and high interclass similarity among the character images. Unlike machine printed texts, handwritten texts vary in terms of the size and style of the characters written. These dissimilarities pose a challenge for HCR systems as the same character can be written in many different ways by different individuals. Also, there are variations in the characters written by an individual depending on his/her state of mind at the time of writing or other factors such as the physical position of the person, the instrument used for writing, etc.
- Shape similarity: Some characters have similar shapes. These characters when written by hand are more difficult to distinguish from each other. In Meitei Mayek, character pairs such as ꯀ and ꯈ, ꯆ and ꯖ exhibit similar

shapes. When the complete character set is considered, the task of recognition becomes more difficult as there are more number of similar-shaped characters. Two more examples in Meitei Mayek script are ꯃ and ꯨ, ꯪ and ꯰.

- Noisy images: Handwritten document images of poor quality also make the recognition task challenging. Noise may be in the form of low resolution of scanned documents or low paper quality of documents.

- Script-specific peculiarities: Most Indic scripts have unique characteristics like the presence of shirorekha (header-line), matra, compound characters and modifiers. Such peculiar features tend to produce more errors in the recognition system.

- Skewness of data: The handwritten text may be skewed due to skewness introduced during scanning or due to the skew nature of the individuals' handwritings.

## 1.2 Handwritten character recognition approaches

The main stages involved in the process of off-line handwritten character recognition are: image acquisition, pre-processing, segmentation, feature extraction, classification & recognition and post-processing. Many methodologies have been proposed for each stage. A detailed account of the earlier methods proposed for different stages is provided in the survey conducted by Arica [16]. For Indic scripts, extensive study on the several off-line techniques proposed has been carried out in the works reported by Pal et al. [149, 153]. The techniques adopted for HCR systems can be broadly divided into four categories depending on the features used. They are:

1. Template matching technique
2. Techniques based on handcrafted features
3. Techniques based on deep-learned features
4. Techniques based on combination of handcrafted and deep-learned features

### 1.2.1 Template matching technique

In this technique, image pixels of the input image are compared pixel by pixel with a set of predefined templates which is believed to represent the entire character class [219]. This technique involves superimposing the image onto the templates and finding the degree of similarity between the input image and templates. The input image is then recognized as the template which matches most closely with it. Template matching technique is time-consuming and not efficient for those situations where the images are noisy and it is sensitive to size, style and font of the images.

### 1.2.2 Techniques based on handcrafted features

In this class of techniques, handcrafted features are extracted from the images using different methods by actually analysing the images and their properties. The work involves human intervention such as that of data scientists who try and come up with features which are discriminative enough for recognition to carry out amongst the character classes. The features which are extracted are compared with those of the true class characters. The images are then classified into the class for which the most similar features are found. Most of these features have been manually designed with an aim to overcome specific issues that exist in image recognition tasks such as illumination and scale variations and occlusions. They also depend on the data they are used for and the application they are applied to. The main problem with handcrafted features is that feature engineering is difficult and takes a lot of time. Many handcrafted features have been proposed and analysed for HCR. They can be classified as *Global Transformation and Series Expansion*, *Statistical Representation* and *Geometrical and Topological Representation* [16]. Some recent and popular state-of-the-art handcrafted feature descriptors are Histogram of Oriented Gradients (HOG) [42], Local Binary Pattern (LBP) [146], Scale-Invariant Feature Transform (SIFT) [121], Speeded Up Robust Features (SURF) [25] and Oriented FAST and rotated BRIEF (ORB) [176].

### 1.2.3 Techniques based on deep learned features

The deep learned features-based techniques are the techniques where features extracted are the features learned automatically by deep networks like CNN. A dedicated stage of feature extraction is not required to extract deep features unlike the case with traditional handcrafted features. Features can be extracted from different layers of CNNs. Features extracted from lower/shallower levels are more generalized in nature and features extracted at higher/deeper levels tend to be more specific to the training images. Different features could be extracted from different layers depending on the type of features one is looking for. Features from shallower layers known as low-level features are called local features as the features describe certain properties which are based on the local neighbourhood of an image's surface pixels. Features extracted from higher layers are called global features as the features represent the entire input image and are extracted from local features. These two types of features encode different-level information. High layer features represent the semantic information while low layer features represent the detail information of the image [53].

### 1.2.4 Techniques based on combination of handcrafted and deep-learned features

This class of techniques employs a fusion of handcrafted and deep-learned features. The two types of feature descriptors have their own advantages when it comes to their performance, both in terms of computation cost and feasibility. Handcrafted features can be manually designed and can be used for specific classification task, such as finding a specific feature descriptor which discriminates one character from another highly shape-similar character. They are computationally less costly compared to deep features. They can also be used to train classifiers when the amount of training data available is not large enough to train a deep network. Deep features on the other hand have shown to give state-of-the-art performance at the expense of higher training time and larger amount of training data for training the deep networks.

Works related to the different techniques have been provided in the corresponding individual chapters in the later parts of the thesis.

## 1.3 Historical background and evolution of present HCR systems

HCR is a sub-field of Optical Character Recognition (OCR). With the success of OCR of printed and hand-printed characters, HCR came into being. OCR is a research area which has been extensively studied for more than a half century now. The history of OCR can be traced back to 1900, when the Russian scientist Tyuring attempted to develop an aid for the visually handicapped[126]. One of the pioneering works reported in character recognition is that of Grimsdale et al.[69] in the year 1958. Here, the input pattern is scanned by a flying spot scanner and a digital computer is used to analyze the shape of the input patterns and to extract the basic features. These features are compared with the values stored in the computer and are recognized accordingly. Another notable work was put forward by Murray Eden[54] [55] where it was shown that all characters of Latin script can be formed by 18 strokes, which can be generated from a subset of four strokes. This concept later gave rise to what came to be known as ”analysis-by-synthesis method”.

One of the first OCR systems for printed characters was developed in the 1940s using template matching. During the 1950s, researchers realized that developing templates for OCR of hand-printed characters was a difficult task because of the many variations in handwritten texts. This led to the introduction of techniques based on handcrafted features. The first commercial OCR named IBM 1418 was developed in the early 1960s. Later in 1965 when the technology advanced, the first HCR named *IBM 1287* which could read handwritten numerals was introduced in the New York world fair [135]. Commercial OCRs which could read poor print quality characters and hand-printed characters were developed between 1975-1985. In the beginning of 1990s, artificial intelligence (AI) techniques were successfully combined with image processing and pattern recognition methods. Efficient methodologies such as neural networks (NNs), Hidden Markov models (HMMs), fuzzy set reasoning, and natural language processing started gaining momentum and aided in the success of more powerful computers [16]. During the late 1980s and 1990s, more number of commercial OCRs were introduced in the market which could deal with more sophisticated text documents such as multi-font and postal address etc. Integration of OCR with spell-checkers and speech output became available in the market. The problem of HCR was still a challenging one during this time period and remained to be an open research area [148].

In the late 1990s and the current decade, research in HCR saw a boom with one of the earliest and successful works using deep networks carried out by LeCun et al. [111]. They used backpropagation and gradient-based learning to train a Neocognitron-like architecture for the purpose of document recognition. Researchers also started working on different learning approaches such as Support Vector Machine (SVM), K Nearest Neighbour (KNN), Random Forest (RF), Decision Tree (DT), Neural Networks (NNs), etc. In the early 2000s, researchers proposed a number of handcrafted feature descriptors such as HOG, SIFT, SURF etc which gave state-of-the-art results on a number of HCR systems during that time. Combination of machine learning methods and handcrafted feature descriptors could increase the accuracy of the OCR systems especially HCR. Then came the deep learning paradigm which can be considered a milestone in the OCR research community. Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) networks are some of the most successful deep learning architectures [112]. In the year 2010, the famous ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [177] was introduced to the research community which further proved the popularity and success of CNNs for image classification tasks. Many state-of-the-art CNN models were proposed as part of this competition. Like other image classification tasks, CNNs and techniques based on transfer learning using pre-trained models have been the state-of-the-art performers of present day HCR systems.

## 1.4 Meitei Mayek

Meitei Mayek is one of the few scripts belonging to North-eastern part of the Indian subcontinent. It is used for writing a Tibeto-Burman language [131] called Manipuri which has been recognised as one of the 22 scheduled languages under the Indian Constitution. The first use of the script dates back to the 11th century and continued to be in use till the 17th century. In the early 18th century, a Brahmin missionary introduced Hinduism/Vaishnavism in Manipur which was then a princely state. The king of Manipur at that time, Pamheiba ordered to burn down all the existing scriptures written in Meitei Mayek under the influence of his newfound love of Hinduism. This fateful day in the winter of 1729 became to be known as the ”Puya Meithaba”- whose literal translation would be ”The Burning of Puya.” Puya is a repository of ancient manuscripts written in Meitei Mayek. From these facts, one could only wish that there was a way to store the scriptures and handwritten texts in some form not destroyable by fire. Many scholars, historians and social scientists believe that this event had led to the dismantlement of culture, identity, rituals, language, and the traditional religious system of the people of Manipur.

### 1.4.1 Origin and history of Meitei Mayek

The exact date of origin of the script is unknown due to the lack of evidence of its sources. There exist different viewpoints of different researchers regarding the same. Mythologically, it is believed that the existence of the script was seen as early as in the late 4th or early 3rd millennium B.C. According to Professor Kalidas Nag, it was long before the Asoka period that the script was developed [1] [200]. There is another belief that the script was created in the late Haya Chak (second age) before casteism existed among the Hindus [118].

Many stone inscriptions were found in different parts of Manipur such as Ngaprum Chingjin, Khoibu, Leishangkhong, etc. The epigraphical studies of these inscriptions affirm that the ancient civilization of the state (Manipur) had its own literature.[2] [200]. Manipuri is also believed to be one of the oldest languages in India with a rich ancient and medieval literature. The Meitei Mayek script was not only used during the British reign, but also much earlier during the time of the sovereign kings of Manipur. [3].

Apart from the controversial history of the script, there is another controversy regarding this peculiar script. It is in regards to the number of letters to be used for the script. There are four Schools of thought who propose four different sets of letters. The first one states that there should be 18 letters claiming that those are the only original Meitei Mayek script. They believe that the origin of the letters should be directly related to the 18 parts of the human body.[4]. The Second one says that only 27 indigenous archaic letters were originally used in the Puya, so the number of letters of the Meitei Mayek script should be 27. [5]. The third one argues that 35 letters were used during the reign of Maharaja Chandrakriti Singh of Manipur [6]. There was a traditional belief that these letters were created through devine characters of Gods and were used to record the events in the ”Puyas” [7]. The fourth School maintains that according to the epigraphical records found in the form of coins, the script is one of the oldest scripts in the

[1] Thambal Angou Singh: Meitei Mayek, Imphal 1947, P.2

[2] Dr.P.Gunindro: Origin and Development of Meetei Script (A paper submitted in the seminar 1988 P.I.)

[3] A memorandum submitted to the hon'ble chief Minister by twenty two literary & cultural organisation. Date 4/1/91, P.4.

[4] W. Tomchou Singh: Meitei Mayekki Lairik Ahanba, New DElhi (Okhla Industrial Area), 1992, P.P 13-18

[5] O.J.M. & M.B.S. 2000. Mapi Lairik Ahanba Wakhal. Imphal: The North-East Meetei Mayek Academy (NEMMA)

[6] ”Meitei Yelhou Lairik”, Pub. by Birodha Janani Sabha, Under the Royal order of Sir Churachand K.C.S.I. & C.B.E., 1934. P.I.

[7] Meitei Mayek Sub-Committee Report, Annexure-II, P.23, 1992-93

human civilization. Their claim is mainly based on the original sounding system of the script which was divided into vowels known as ”Thingdaba Mayek” and consonants known as ”Phongdaba Mayek” [8].

### 1.4.2 Present day Meitei Mayek

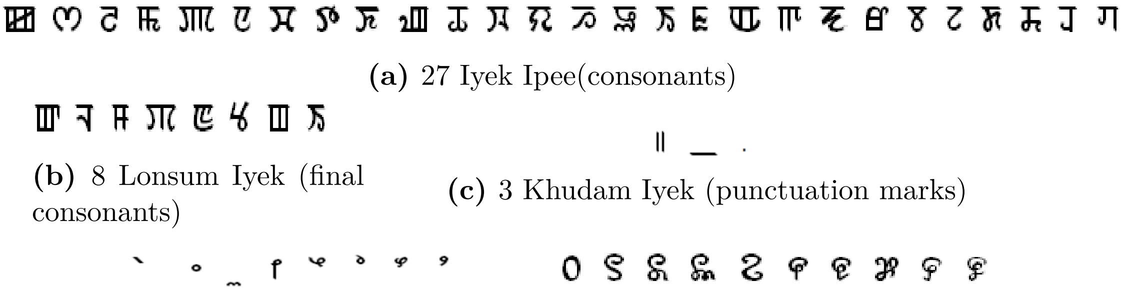
Realising the importance of the script, a lot of attempts were made in different forms during the period from 1930 to 1980 by the people of Manipur to reinforce the script to use. During this period, there were debates on the number of letter-forms to be included in the character set, whether it should be 36 or 27. Finally in the year 1980, the Government of Manipur came to the decision to accept 27 as the final number of letters (Manipur Gazette, No. 33, 1980). It can be said that the script came back to life as it was brought back to use after many centuries. And it was introduced for the first time as a part of academic curriculum in the academic session of 2005-2006 [98].

Since Meitei Mayek is a script which got revived after centuries of disappearance, a bigger section of the population of Manipur do not know how to read and write the script. So therefore, the present time is a transition phase from Bangla to Meitei Mayek. This shows the necessity to have a transliteration system to transliterate handwritten texts written in Meitei Mayek to Bangla and vice-versa. This will assist to bridge the communication gap, in the context of handwritten documents, between the mass of people who do not know Meitei Mayek and those who know Meitei Mayek. Since the adoption of Meitei Mayek in the academic front, many natural scenes such as sign boards of shops, institutes, hospitals, vehicle number plates, etc. use Meitei Mayek. For a person who only knows Bangla script, to read such natural scenes, a transliteration system which converts it to Bangla is required. And the first step to achieve this task is to convert the written Meitei Mayek text into a format that the system can process.

There are a total of 56 letters in the character set of Meitei Mayek. This includes 27 consonants (Iyek Ipee), 8 final consonants (Lonsum Iyek), 8 vowels (Cheitap Iyek), 3 punctuation marks (Khudam Iyek) and 10 numerals (Cheising Iyek) (Figure 1-1). Out of the 27 consonants, 18 are original and 9 are evolved alphabets known as lom Iyeks. The 18 original alphabets are named after the names of the parts of the body [123]. For instance, the name of the first consonant ꯀ is *kok* which means ”Head” in Manipuri, ꯁ is called *sam* meaning ”Hair” and

[8] Moirangthem Munan: Meetei Mayek Ahanba Lairik, Pub. 1966, P.P. 1-2

so on. The 9 evolved alphabets are derived from 9 of the original 18 alphabets. For example, ꯒ (gok) evolved from ꯀ (kok), ꯔ (rai) from ꯂ (lai). The ancient sacred books called ’Puyas’ and many other ancient manuscripts were written using only these 18 original alphabets and some vowel derivatives. Even in the modern dictionary, every word begins with one of the 18 original alphabets. And the 9 evolved alphabets are present at the end or between the first and the last alphabets of any word. Any other words beginning with a letter other than these 18 original alphabets is a loanword and they are limited in number. There are also no clusters or compound characters in Meitei Mayek. Two consonants are merged with the used of a third letter called *apun* which is a line drawn at the bottom of the two consonants. The shapes of the 10 numeral figures of Meitei Mayek are derived from the shape of the human embryo. Each numeral represents the shape of the embryo that undergoes a change every month during its development of 10 months.



**(a)** 27 Iyek Ipee(consonants)

**(b)** 8 Lonsum Iyek (final consonants)

**(c)** 3 Khudam Iyek (punctuation marks)

**(d)** 8 Cheitap Iyek (vowels)

**(e)** 10 Cheising Iyek (numerals)

**Figure 1-1:** Characters of Meitei Mayek script

Meitei Mayek follows the Abugida type of writing system and is written from left to right direction. The script is case-insensitive. It was added to the Unicode Standard in October, 2009 with the release of version 5.2 [9]. Meitei Mayek is a tonal language and there exists a character called *lum* (.) which is used as a tone marker. This character is used only as a sign of intonation, it is not used in writing. It was reported that a Meitei Mayek text corpus of 2,65,540 characters does not have a single occurrence of . [3]. It was also observed during the data collection process that the use of . is zero in all the data samples that were collected. Therefore, the number of characters used for writing Manipuri language is 55 which is also the number of characters taken into consideration in the present research.

[9] https://unicode.org/charts/PDF/UABC0.pdf

## 1.5 Existing works in HCR of Meitei Mayek

Tangkeshwar et al. [211] was one of the first researchers to work on handwritten character recognition of Meitei Mayek. In their work, the image is first binarized after which the character pattern is segmented using heuristic segmentation technique. KL-divergence technique and a neocognitron simulator was used for feature extraction and recognition of segmented characters and achieved an accuracy of 90%. In another work by them [213], probabilistic and fuzzy features were used with neural network as classifier. It was found that the accuracy achieved was 85.92% for probabilistic features alone, 88.14% for fuzzy features alone and 90.3% for the hybrid features.

HCR for handwritten Meitei Mayek numerals [107] and alphabets [108] using multilayer feed forward neural network with backpropagation learning was reported by Laishram et al. The overall accuracy achieved was 85% for the numerals. In the latter work, the focus is given on the segmentation of character from a scanned whole document. Segmentation of lines and words is done using histogram-based algorithm and segmentation of characters is done using a connected component method. The accuracy was measured in two ways. The first one was where the recognition achieved was with respect to the output of a segmentation stage given to the neural network. The accuracy is 80% in this case. The second case scenario was where the input was given to the neural network itself. The accuracy reported in this case is 85%. They concluded that accuracy is affected by the segmentation stage because as the images are dilated, connected component analysis technique tends to consider adjacent characters as a single character during segmentation. Maring and Dhir [127] achieved an accuracy of 89.58% for recognition of Meitei Mayek numerals using Gabor filter-based technique and SVM classifier.

Kumar and Kalita [99] used various feature extraction techniques such as zone-based diagonal, background directional distribution (BDD), Histogram of Oriented Gradients (HOG) and projection histograms with SVM classifier for recognition of Meitei Mayek numerals. They reported that the HOG feature is more appropriate as compared to others and highest recognition of 95.16% is obtained for the combination feature set consisting of HOG, histogram, BDD, and diagonal features. Kumar et al. [101] also achieved accuracy of 94%, 92% and 98% using distance profile features (DPF) and background directional features (BDF) and hybrid (DPF+BDF) respectively on a dataset of 27 consonants.

Tangkeshwar [201] reported the use of vertical and horizontal projection profiles and connected component analysis methods for segmentation of isolated digits and non-touching characters. Probabilistic features (PF) and fuzzy features (FF) were used with K-L divergence technique and feed forward back propagation neural network (MLPs) for recognition. Recognition using the hybrid feature set (PF+FF) was also performed and it was concluded that it gave better recognition rate than that using PF or FF alone. An algorithm based on the zoning information to recognize a word with isolated and overlapping characters in the Meitei Mayek script was also proposed. It was done so to test the trained neural network for that word only and it was not a generalized algorithm. A hybrid point feature based recognition using Harris corner detector, Gilles feature, Laplacian-of-Gaussian (Log) detector and Harris-Laplacian detector was reported by Kumar et al. [100]. An accuracy of 97.16% was achieved using SVM. Nongmeikapam et al. [143] in their work used HOG features with k-NN classifier and achieved an accuracy of 94.29% on a dataset of 56 classes. In another work by Nongmeikakpam et al. [144], HOG feature descriptors are used to feed CNN and the extracted features from CNN are fed to a KNN classifier. They carried out the work with a total of 5600 samples for 56 character classes and reported an accuracy of 98.71%. Inunganbi et al.[87] reported the use of three-channel image by combining gray channel, gradient magnitude channel and gradient direction channel which is fed to a CNN. They could achieve an accuracy of 98.70% on a 35-class Meitei Mayek dataset. The authors also reported work based on a combination of projection histograms and variations of LBP [86]. An accuracy of 98.16% is reported with KNN classifier for a 35-class dataset. Approaches based on CNN were also carried out by them which achieved an accuracy of 98.86% [85] and 99.02% [88].

## 1.6 Motivation

HCR has been around in the research community for a long time. However, there are still some scripts which are not explored in this aspect. An HCR system developed for a script or multiple scripts might not perform well with certain other scripts. This is because of script-specific features possessed by different scripts. Unless those systems are tested against a dataset of the concerned script, one might not be able to tell the applicability of the systems for that specific script. As far as Meitei Mayek is concerned, there are some works reported in literature. In spite of those works, there are still scopes of improvement to achieve a better recognition accuracy.

The motivation behind the work carried out in this dissertation is the research gaps that are observed from the concerned script point of view. They are:

- No standard publicly available datasets are found. The previous works are carried out on datasets which are created in lab environment. Also, the datasets used in most of the works are subsets of the complete character set and not the complete set.
- Deep learning techniques are less explored for the concerned script. Deep learning techniques especially CNNs have shown remarkable performance in HCR tasks. They could be used as part of the overall system in order to tackle script-specific issues such as identification of characters having high structural similarities.
- Script-specific properties have not been considered so far in literature. Meitei Mayek words can be divided into zones and certain orthographic rules are followed while writing the script. These specific properties of the script could be utilized in the recognition process to have an enhanced recognition accuracy.
- Language modelling can improve the HCR system by incorporating the contextual information to the system. The thesis also seeks to overcome some of the limitations of the method that uses script-specific properties.

With the replacement of Bengali script by Meitei Mayek script to write Manipuri language, the development of such a system will prove beneficial for the government as well as the people in general. Since the present time is a transition phase from Bangla to Meitei Mayek, development of a robust OCR is very important which in turn is pivotal for development of other systems such as machine transliteration system, especially one which translates Meitei Mayek script to Bangla script and vice-versa.

## 1.7 Objective

The objective of this dissertation is to develop methods for off-line HCR of Meitei Mayek characters. Through this Ph.D research work, four goals have been established to achieve the objective:

- to develop a dataset of complete character set of Meitei Mayek script
- to develop a Convolutional Neural Network (CNN)-based recognition for the developed dataset
- to propose a methodology for multilevel recognition of Meitei Mayek handwritten characters using fusion of feature strategy.
- to identify and utilize script-specific properties for achieving enhanced recognition accuracy
- to incorporate language model (LM) in the recognition system for a better recognition accuracy

## 1.8 Dissertation contributions

The main contributions of the thesis can be divided into four parts. The contributions have been briefly outlined in the following subsections:

### 1.8.1 Creation of a Meitei Mayek handwritten character dataset (TUMMHCD) and performance analysis of existing features with state-of-the-art classifiers

Study on HCR of any script requires availability of a dataset of the concerned script. At the time of start of this thesis work (2017), there were no publicly available datasets of Meitei Mayek in literature. As part of the present work, a dataset of handwritten Meitei Mayek characters has been developed. The name given to the dataset is Tezpur University Meitei Mayek Handwritten Character Dataset (TUMMHCD). It consists of a total of 85,124 character images. In each class, the images are randomly divided into training set (85%) and test set (15%). The training set and test set have a total of 72,330 and 12,794 images respectively. Many techniques have been proposed for HCR over the decades by researchers. The performance of two popular handcrafted features and raw image pixel intensity values are evaluated against the developed dataset using 4 state-of-the-art classifiers.

### 1.8.2 Development of a Convolutional Neural Network (CNN)-based recognition for the developed dataset.

With the advent of deep neural networks, many computer vision tasks have seen state-of-the-art performances. CNN has shown its superiority as a recognition tool, and hence it is the obvious choice of many researchers. The success of CNN can be attributed to its unique properties such as it

- can an be used as both feature extractor and classifier
- takes advantage of local spatial coherence of the pixels
- uses shared parameters and hence less number of trainable parameters to deal with
- is robust to noise and is shift-invariant

However, one major challenge in developing a CNN model is that there are no rules of thumb to follow. The adopted architecture of a CNN is really a design choice which depends on the concerned designer(s). That said, there are certain hyperparameters that need to be tuned in order to achieve the best performing model. The performance of a CNN model also depends on the dataset being evaluated against it. In this goal of the research, a CNN model is built from scratch by tuning the hyperparameters and thereby setting a benchmark on TUMMHCD. The performance of the proposed CNN model is compared against five state-of-the-art CNN models viz. InceptionV3, ResNet50V2, DenseNet121, MobileNetV2, and EfficientNetB3 in terms of training time, testing time, validation and test accuracies. The experimental result proves that the proposed model show superior performance.

### 1.8.3 A multilevel recognition of Meitei Mayek handwritten characters using fusion of feature strategy.

During the performance analysis of the works carried out in the previous two goals, it is observed that there are certain characters which share common shapes and hence are more difficult to distinguish from each other. The occurrence of these characters leads to a lower recognition accuracy of the system. To address this problem, a multilevel methodology with fusion of features is proposed. It is called *multilevel* because it has two levels of recognition and the second level of recognition adopts a fusion of multiple features, both handcrafted and deep learned features. The fusion of handcrafted and deep features have been adopted in many pattern recognition and computer vision problems but not many works are found in HCR. The multilevel approach is also adopted in many HCR problems which deals with different scripts. To the best of our knowledge, no methodologies in HCR have been proposed so far for the combination of multilevel and fusion of feature in literature.

### 1.8.4 Identification of script-specific properties for achieving enhanced recognition accuracy.

In any script, there are some characters which get misrecognised when fed to the recognizer. In order to recognize such similarly shaped characters, there is a need for additional type of information which are more discriminative in order to distinguish among these characters. In most works reported in literature on HCR with confusing character pairs, it is observed that feature vectors considered are either the traditional hand-crafted features or features extracted from neural networks. The works lack in the area of incorporating the peculiarities that exist in a specific script. This goal of the research aims at exploring the script-specific properties that exist in Meitei Mayek script by finding out the zonal information and orthographic rules. These rules which are specific to the script, once found out can be of significant importance in enhancing the recognition accuracy of a particular script. Moreover, for those scripts which follow a strict set of rules while writing them, the rules can hardly go wrong while incorporating it in the recognition system. Another advantage of the present approach is that the second stage recognition only takes a comparison function to make decision unlike the previous works where feature extraction and classification takes place in all the stages. This reduces the computational cost in the present work.

### 1.8.5 Incorporation of language model in the recognition system.

The work described in the previous goal achieves an improved recognition accuracy in HCR of Meitei Mayek. It, however, is script-specific and the method cannot be generalized to other scripts. Identifying this limitation, an approach to incorporate language model with CNN is introduced. The method leverages the power of visual understanding of images through CNN and the power of contextual understanding of text through an LSTM-based LM. LM works as a post-processing step. The recognised text output of the CNN is fed as the input to the LM. The LM then generates the conditional probabilities (CoPs) which is the probability distribution of the next character token based on the prior context. The proposed methodology not only considers CoPs but it also takes into account the decision made by the CNN by combining the class probabilities generated by the CNN and the CoPs to give the final output text.

## 1.9 Organisation of the thesis

The thesis is organised as follows:

### Chapter 1: Introduction

This chapter presents an introduction to HCR and its applications. A background literature about OCR and the works thereafter is also provided. It also gives a brief account of the concerned script and the motivation behind taking up this research work. This chapter also lays out the objectives and contributions of the research work.

### Chapter 2: Dataset creation and performance analysis of state-of-the-art classifiers with existing features

This chapter deals with the creation of a complete character set of Meitei Mayek handwritten character dataset named Tezpur University Meitei Mayek Handwritten Character Dataset (TUMMHCD). A highlight of the existing publicly available datasets is also given. Performance evaluation of two popular handcrafted features and raw image pixel intensity values with four state-of-the-art classifiers is presented in this chapter.

### Chapter 3: CNN-based recognition of TUMMHCD

This chapter first briefs the related works. A performance evaluation of five state-of-the-art CNN models against TUMMHCD is provided. A CNN model built from scratch for recognition of TUMMHCD is proposed. The performance of the proposed CNN model is then compared with those of the state-of-the-art CNN models.

### Chapter 4: A multilevel recognition of TUMMHCD using fusion of features strategy

This chapter gives an introduction to different methods of feature fusion found in literature and presents a brief account of related works. It then describes the proposed multilevel approach using fusion of handcrafted features and deep features for recognition of TUMMHCD. Experimental results and discussion are provided.

### Chapter 5: Zone and rule assisted recognition of Meitei Mayek handwritten characters

This chapter first introduces multistage recognition of handwritten characters and gives a brief account of the related works. It then describes a multistage method which takes advantage of script-specific properties viz. zonal information and orthographic rules for recognition of Meitei Mayek script.

### Chapter 6: Using LSTM language model with CNN for handwritten character recognition

This chapter first highlights the need for LM in enhancing the performance of an HCR system. It also identifies the bottleneck facing the HCR of Meitei Mayek w.r.t. incorporating language modelling. It then proposes a methodology of combining the strengths of CNN and LM for achieving an enhanced recognition accuracy.

### Chapter 7: Conclusion & future direction

This chapter concludes the thesis with a summary of works done and lists possible works that can be carried out in future in this area.