I enjoy doing things that involve research because it's part of what I enjoy about acting.

- Michael Sheen

CHAPTER 3

MEETEI MAYEK SCRIPT AND DATASET

<table>
<thead>
<tr>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
</tr>
<tr>
<td>3.2</td>
</tr>
<tr>
<td>3.3</td>
</tr>
<tr>
<td>3.4</td>
</tr>
<tr>
<td>Introduction</td>
</tr>
<tr>
<td>Meetei Mayek Script</td>
</tr>
<tr>
<td>Dataset Preparation</td>
</tr>
<tr>
<td>Conclusion</td>
</tr>
</tbody>
</table>

3.1. INTRODUCTION

One of the major challenges to perform optical character recognition research in Indian regional script is the non-availability of standard dataset. In Meetei Mayek also no standard dataset is publicly available. A dataset is prepared for carrying out the research work on Optical Character Recognition of Meetei Mayek Script. Various possible variations of handwriting is accommodated so that the experimental results match real life writing.

This chapter gives an introduction about Meetei Mayek Script, its origin, an overview of how the dataset is prepared followed by various pre-processing operations.

3.2. MEETEI MAYEK SCRIPT

North-East India consists of eight states called eight sisters, the states of Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim and Tripura. The North-Eastern region of India can readily be considered as miniature India. Each of the seven states has a large number of different linguistic groups. These groups, from different linguistic domain remained confined within a small and
limited logical domain for decades together. This may be a cause of a big social barrier, against the social homogeneity and national integrity among these ethnic groups. Except Assamese and Bengali, most of the languages used by the native speakers are of Sino-Tibetan and Austro-Asiatic origin. Meetei is one of the major northeast Indian regional languages [86]. It is primarily spoken by majority of the native speaker in the state of Manipur. It is the mother tongue of the ethnic group Meitei. Linguistic affinity Manipuri belongs to the Sino-Tibetan family of languages. Figure 3.1 shows a page from Meetei Ancient book of knowledge [86] and Figure 3.2 shows a schematic diagram of the Sino-Tibetan language family.

Meetei Mayek script is used for writing Meitei (Manipuri) language. The script has a long history behind it. The Meetei Mayek script was used in earlier days for writing Meitei till 18th century, most of the historical documents found are written using Meetei Mayek script [87]. However, there was a mass destruction of documents written in Meetei Mayek in 18th century, and Bengali script was adopted for writing Meetei thereafter. The script almost became wiped out as an outcome of a mass burning of all books written in Meeteilon ordered by Ningthau Pamheiba who was the ruler of Manipur in the 18th century.

Figure 3.1: a page from Meetei Ancient book of knowledge
Figure 3.2: Schematic diagram of the Sino-Tibetan language family
It is really difficult to trace the precise period of the beginning of the Meetei Mayek. The burning of crucial historical credentials or the Puyas of Kangleipak (Manipur) written in Meetei Mayek in the reign of King Pamheiba in the early 18th century, made the effort all the more difficult [88]. The earliest use of this very script is dated between 11th and 12th centuries AD.

Figure 3.3: Meetei Mayek characters

Good news is that, researchers worked over the script and finally in 1976, they have finalized the characters in the script having 18 characters and added 9 more letters which are derivatives of previous letters and suggested the government to make the script popular in the state of Manipur. In 1983, the education department of Manipur Government start publishing text book written in Meetei Mayek. Meetei Mayek script has 18 characters in original which are called Eeyek Eepee and 9 additional characters incorporated are called Lom Eeyek. Figure 3.3 shows the characters of Meetei Mayek. The most interesting fact about the script is that some of the letters used in Meetei Mayek refers to human body part.

3.3 DATASET PREPARATION

Dataset used in our experimental work for recognition of Meetei Mayek characters consists of 12960 samples. For this study 27 basic characters of Meetei Mayek alphabet are considered.
3.3.1 DATASET COLLECTION

These samples are collected from writers of different profiles, age and profession. Each writer has contributed 10 samples of each of 27 basic characters, thus forming a dataset of 12960 samples. Writers who contributed to these samples were having different writing styles. In some characters distortions are also introduced and amount of such distortion depends on quality of pen ink used to write characters and speed of the writer to write the characters. Standard data collection sheet of A4 size was provided to the contributors for writing the characters of Meetei Mayek script. Figure 3.4 shows data sample contributor’s distribution.

![Statistics of Contributors](image)

**Figure 3.4:** Data sample contributor’s distribution
Each of the characters was written ten times in a row. Separate characters are written in separate row. Writers were given the choice to use any pen with ink of any color; no constraint was imposed on them over writing. They have contributed their natural writing. Printed characters were given on a separate page for reference.
per contributor was collected. Most of the contributors are from S.S.N school, Cotton University and Gauhati University.

3.3.2 DIGITIZATION AND PRE-PROCESSING

SCANNING

The documents collected are digitized by using a scanner with dpi 300. The scanned images are numbered and stored in .tiff format. Noise removal from the images is performed. Line and character segmentation algorithm are used to get isolated character images.

SEGMENTATION AND NOISE REMOVAL

In the present study the dataset is collected on A4 size plain paper and the individual characters are segmented using algorithm 3.1 and 3.2.

Algorithm 3.1: Line Segmentation

Input: A4 size scan document

Output: Segmented line of characters

1. Convert the image to binary
2. Base lines of the text are determined for the first \(1/8\)th width of the entire text document.
3. The baseline is the row with minimum number of pixels.
4. The window for which the baselines are obtained is then moved \(1/8\)th width towards right.
5. The process continues till all the baselines of the entire text document are found. The text lines are then separated along these baselines.
6. End
The output of the Algorithm 3.1 is taken as input in Algorithm 3.2 for character segmentation.

**Algorithm 3.2: Character segmentation**

**Input**: Segmented line  
**Output**: Isolated character with reduced noise

1. Erosion with circular structural element is done  
2. Connected Component detection is performed  
3. Removal of objects with less than 20 pixels is done  
4. Dilate and label the isolated images  
5. End

The individual characters are segmented using above mentioned technique.

**PRE-PROCESSING**

Pre-processing techniques applied on these isolated characters, so as to make these samples easy to extract efficient features.

RGB to Gray Conversion: Entire images are RGB images, are firstly converted to Grayscale images. Figure 3.6 shows RGB and Gray Scale image of Meetei Character ‘kok’.

- **Gray to Binary Conversion and negative of images**: Gray scale image is then converted to binary image using Otsu’s threshold method. Binarized image of character after conversion from Gray Scale image is shown in Figure 3.7 The negative of these binary images are then extracted.
Figure 3.7: (a) Gray Scale image of Meetei Character ‘kok’ (b) Binarized inverted image of Meetei Character ‘kok’

**Median Filtration:** Median filtration on the Binarized images is applied. Each output pixel is changed to the median value in the 3-by-3 neighborhood around the corresponding pixel in the input image.

### 3.3.3. DATASET GENERATION

![Sample of Dataset1](image)

Figure 3.8: Sample of Dataset1

Literature review reveals that Meetei Mayek character recognition has been carried out on non-skewed dataset. In the present study we have developed two variations of data on Meetei Mayed script namely dataset1 and dataset2. Dataset1 consist of 12960 images of isolated characters. These characters are preprocessed and normalized to 30x30 sizes as depicted in Figure 3.8.
Dataset2 has been developed from Dataset1. As the real world data may be skewed by any amount, we have randomly rotated the images of Dataset1 by any random amount to create Dataset2.

**Figure 3.9:** Sample of Dataset2

**Figure 3.10:** Characters with class label
This dataset is of great help to analyze how various features behave as characters are skewed at various angles. Figure 3.9 shows samples of Dataset2. Each character in both the dataset have been assigned a class label for classification purpose. Figure 3.10 shows Meetei Mayek characters with their corresponding class label.

3.4. CONCLUSION

This chapter is mainly concerned about the dataset used for carrying out the experiment. It discusses and explains how difficult it is to find benchmark dataset for regional scripts. Starting with Schematic diagram of the Sino-Tibetan language family, a little introduction to Meetei Mayek script is carried out. Data Collection strategy and pre-processing is explained in detail. Segmentation and noise removal techniques are also explained in brief here.