Manipuri Meitei Mayek Handwritten Character Dataset

The sample dataset consists of 5 sample images for each of the 37 classes (10 numerals and 27 consonants).

The *Manipuri Meitei Mayek Handwritten Character Dataset* contains a total of 60285 character images. The distribution of samples over the classes is uneven. There are 37 folders signifying 37 character classes. Description of each class is shown in Table 1 (Numerals) and Table 2 (Consonants).

The collection of data samples was carried out in two phases. The first phase consists of distributing a tabular form and asking people to write the characters five times each. Filled-in forms were collected from different individuals in the age group 12-23 years. The second phase was the collection of handwritten sheets such as answer sheets and classroom notes from students in the same age group. A total of 279 such pages written by 279 different individuals were collected. In total characters from around 500 individuals have been considered for creation of the dataset. The data samples were collected from schools and colleges in different parts of Imphal.

Table 1: Description of numeral classes

Folder Name	**Character Name**	**Corresponding English Name**	**Total no. of images**	**Folder Name**	**Character Name**	**Corresponding English Name**	**Total no. of images**
001_ama	Ama	One	2014	006_taruk	Taruk	Six	1830
002_ani	Ani	Two	1931	007_taret	Taret	Seven	1811
003_ahum	Ahum	Three	1841	008_nipal	Nipal	Eight	1846
004_mari	Mari	Four	1874	009_mapal	Mapal	Nine	1781
005_manga	Manga	Five	1861	010_phun	Phun	Zero	1958

Table 2: Description of consonant classes

Folder Name	**Character Name**	**Total no. of images**	**Folder Name**	**Character Name**	**Total no. of images**	**Folder Name**	**Character Name**	**Total no. of images**
011_kok	Kok	1516	020_ngou	Ngou	1520	029_gok	Gok	1489
012_sam	Sam	1526	021_thou	Thou	1546	030_jham	Jham	1732
013_lai	Lai	1536	022_wai	Wai	1370	031_rai	Rai	1523
014_mit	Mit	1531	023_yang	Yang	1560	032_ba	Ba	1515
015_pa	Pa	1505	024_huk	Huk	1499	033_jil	Jil	1536
016_na	Na	1557	025_un	Un	1536	033_dil	Dil	1533
017_chil	Chil	1548	026_ee	Ee	1506	035_ghou	Ghou	1692
018_til	Til	1554	027_pham	Pham	1438	036_dhou	Dhou	1617
019_khou	Khou	1515	028_atia	Atia	1515	037_bham	Bham	1623

Each character is cropped and their bounding box is found out. Characters inside the bounding box are then size normalized to fit in a box of 24x24 pixels. The images are grayscale images and are saved in TIFF format.

Nomenclature: The image files are named in the following format:

File name: mmhc*i*_*j*.tif where
i: the folder number in the dataset (Eg. 1 for first folder, 2 for second folder and so on)
j: the image number in the folder
tif: Image format

Example: mmhc1_3 is the name of the third image file in the first folder.