Dataset Documentation

Raphael Elvis Rozario ID: 150104121
K M Arefeen Sultan ID: 150104111
Rakib Hossain Ayon ID: 150104120
Radwan Rahman ID: 150104123

April 2019

1 Introduction

In this dataset[1], symbols used in both English and Kannada are available. In the English language, Latin script (excluding accents) and Hindu-Arabic numerals are used. For simplicity it is called the "English" characters set. Dataset consists of:

2 Dataset Description

The detailed description about the dataset is stated below.

- 64 classes ($0-9$, $A-Z$, $a-z$)
- 7705 characters obtained from natural images
- 3410 hand drawn characters using a tablet PC
- 62992 synthesised characters from computer fonts

This gives a total of over 74K images (which explains the name of the dataset). We only uses 64 classes ($0-9$, $A-Z$, $a-z$) for our project. Each file has a data structure "list" with these elements:

1. ALLlabels: class label for each sample
2. ALLnames: sub-directory and name of the image for each sample
3. classlabels: set of labels (classes) in this dataset, coded numerically, e.g. $10 = A$, $11 = B$, ..., $64 = z$
4. classnames: strings of the directories where samples of each class are stored
5. NUMclasses: total number of classes in this dataset
6. TRNind: indexes of the training samples. If 20 splits are used, this is a matrix of N train samples X 20
7. TSTind: indexes of the test samples. If 20 splits are used, this is a matrix of N test samples $X20$
8. VALind: indexes of the validation samples. If 20 splits are used, this is a matrix of N validation samples $X20$
9. TXNind: indexes of the texton samples, i.e., samples used to build the vocabulary with the bag-of-visual-words method. If 20 splits are used, this is a matrix of N texton samples $X20$

References

[1] de Campos, T. E.Babu, B. R. Chars74k dataset, 2009. http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/.