Telugu Optical Character Recognition

CS 725

Muppirala Viswa Virinchi (130070038)
Sai Santhosh Kota (130070042)
Sashank Reddy Gujjula (130070046)
Ravi Kishore Boggarapu (130070048)

April 1, 2017

Table of contents

1. Problem Statement
2. Dataset
3. Approaches
4. Progress
5. Future work

Problem Statement

Motivation

- Telugu OCR it is difficult problem as a single character can be formed by a single vowel or single consonant or it can be compound character consisting of a combination of vowels and consonants.
- Different permutations of the 16 vowels and 37 consonants give rise to approximately 500 unique glyphs.

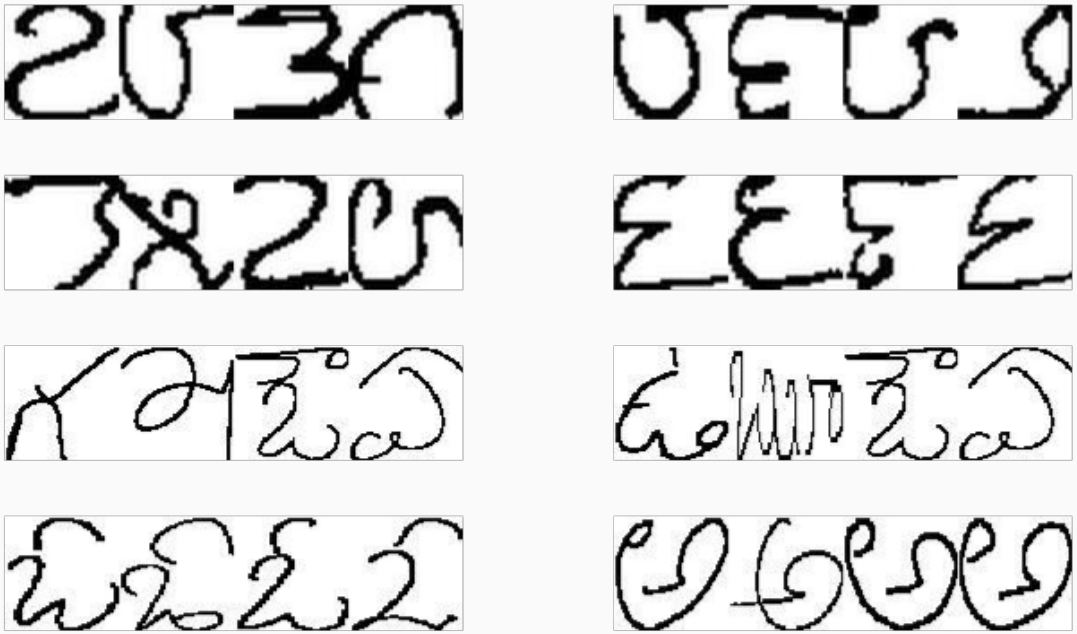
Formulation

- We use techniques like Artificial Neural Networks, Convolutional Neural Network, K-Mean clustering and SVM for recognition
- We take the segmented handwritten telugu numerical data(300*10 numbers) and character data(290*165 characters) and try to formulate a scheme for recognition.
- The class of the image is given by the file name(numbered).

Dataset

Telugu Numerical and Character dataset

Handwritten character data-set was obtained from “HPL Isolated Handwritten Telugu Character Dataset”. Numerical dataset was obtained from ”CMATERdb: The pattern recognition database repository”



Approaches

Approach using ANN, CNN

- Using python libraries: TensorFlow, Keras
- Hyper-parameters of ANN and CNN such as number of layers, filter-dimension, number of filters and stride length are tuned by cross-validation

Approach using K-means clustering and SVM

- Feature Extraction:
  1. Each image is divided into overlapping pieces
  2. K-means clustering is used on these pieces to obtain K centroids which act as features
- Support Vector Machines with appropriate kernel can be used for classification

Progress

Our progress

- All images in the numerical database were of size 32x32. We implemented ANN and CNN on the numerical database using Keras library. We obtained an accuracy of 90.78% using ANN and 91.73% using CNN. Improvements are due.
- ANN architecture:
  - Dense 1024 relu
  - Dense 10 with softmax
  - cross entropy loss function

Our Progress

- CNN architecture:
  - Convolution2D 5x5 30 relu
  - MaxPooling2D 2x2
  - Convolution2D 3x3 15 relu
  - MaxPooling2D 2x2
  - Dense 128 relu
  - Dense 50 relu
  - Dense 10 with softmax
  - cross entropy loss function

Our Progress

- The images in the character database were large and of different sizes
- Dilation followed by resizing(without changing the aspect ratio) to maximum length of 64 was performed on the images
- The images were then padded with white to get 64x64 images

Future work

- K-Means clustering and SVM are yet to be implemented
- Installing CUDA on an NVIDIA machine to speed up the learning process
- Improving Telugu character recognition rate