Machine Learning Engineer Nanodegree

Capstone Proposal

Harathi Surya Patchipala
October 23, 2017

Proposal

HANDWRITTEN TELUGU CHARACTER RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS

Domain Background

Neural Networks are recently being used in various kinds of pattern recognition. Handwritings of different persons are different; therefore, it is very difficult to recognize the handwritten characters. Handwritten Character recognition is an area of pattern recognition that has become the subject of research during the last some decades. Neural network is playing an important role in handwritten character recognition. Handwriting recognition is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch screens and other devices. It can be online or offline. In this context, online recognition involves conversion of digital pen-tip movements into a list of coordinates, used as input for the classification system whereas offline recognition uses images of characters as input.

Although a lot of work has been reported for handwriting recognition in English and Asian languages such as Japanese, Chinese etc., and very few attempts on Indian languages like Hindi, Tamil, Telugu, Kannada etc. In this paper, I am developing handwritten character recognition algorithm for Telugu [South Indian language] with high recognition accuracy and minimum training and classification time.

Some of the earlier works apply shallow learning with hand-designed features on both online and offline datasets. Examples of hand-designed features include pixel densities over regions of image, character curvature, dimensions, and number of horizontal and vertical lines. Shanthi et al. [4] use pixel densities over different zones of the image as features for an SVM classifier. Their system achieved a recognition rate of 82.04% on a handwritten Tamil character database.

Problem Statement

Handwriting recognition has been one of the active and challenging research areas in the field of image processing and pattern recognition. Since 1929, number of character recognition

systems have been developed and are used for even commercial purpose also. Several applications including mail sorting, bank cheques processing, reading aid for blind, document reading and postal address recognition require offline handwriting systems. Working in Postal service need us to decode and deliver something like 30 million handwritten envelopes every single day. The challenges are to do mail-sorting that ensure all those millions of letters reach their destinations.

In this paper an attempt is made to recognize handwritten characters for Telugu language using Convolutional neural networks.

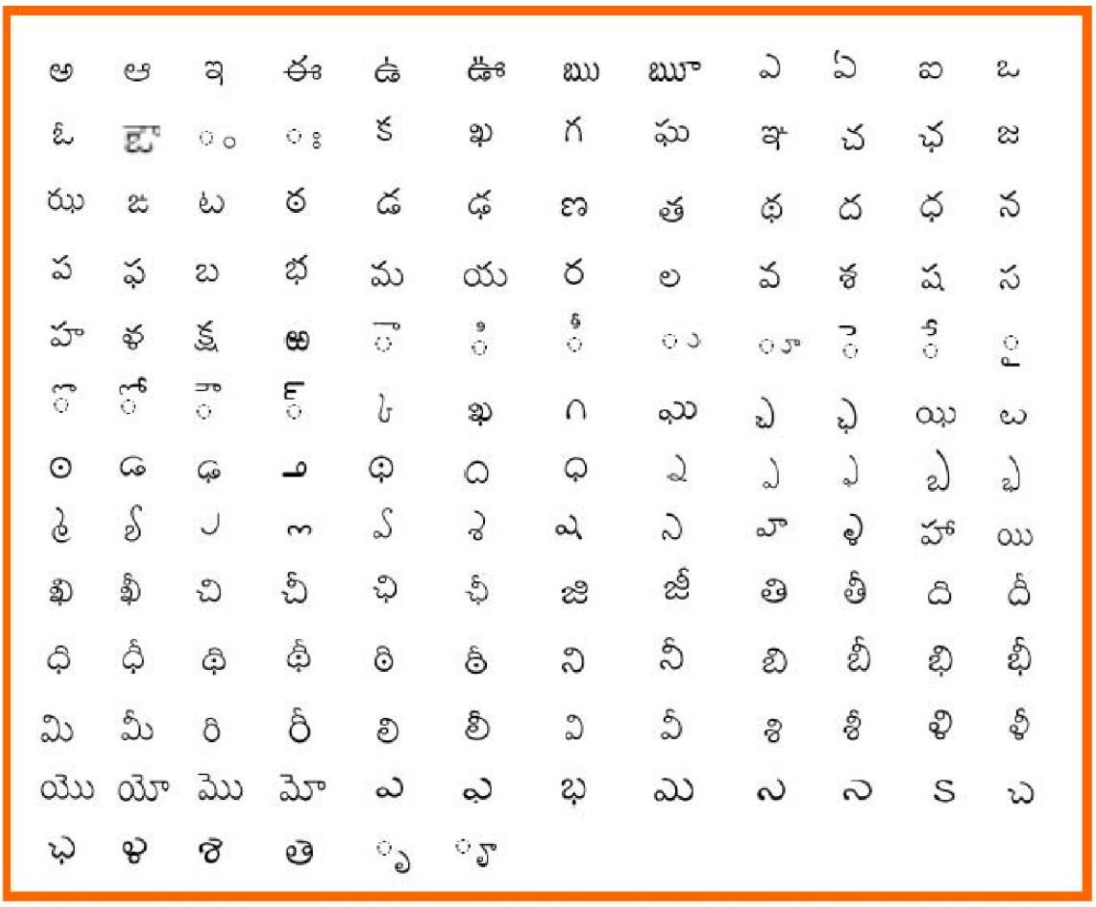
Character recognition complexity varies among different languages due to distinct shapes, strokes and number of characters. [1]Telugu, a South Indian language that ranks third by the number of native speakers in India. Fifteenth in the Ethnologue list of most-spoken languages worldwide and is the most widely spoken Dravidian language in the world. About 800 million people use Telugu as their speaking and writing purpose. Telugu script has 18 vowels and 36 consonants, of which 13 vowels and 35 consonants are in common usage. Of all the Indic scripts, the Telugu script has the largest number of vowels and consonants. Moreover, Telugu contains many similar shaped characters; in some cases a character differ from its similar one with a full-zero (*anusvāra*) (◌ం), half-zero (*arthanusvāra* or *candrabindu*) (◌ఁ) and *visarga* (◌ః) to convey various shades of nasal sounds. That makes difficult to achieve better performance with simple technique as well as hinders to work with Telugu handwritten character recognition.

Datasets and Inputs

Telugu script has 18 vowels and 36 consonants of which 13 vowels and 35 consonants are in common usage. Telugu script is generally non-cursive in style and hence pen-up usually separates the basic graphemes though not always. So, the basic graphemes of the script i.e. independent vowels, consonants, vowel diacritics and consonant modifiers are included in the symbol set. Also included are some consonant-vowel units which cannot be easily segmented. In addition, the symbol set also contains some symbols which do not have linguistic interpretation but have stable pattern across writers and help reduce the total number of symbols to be collected. The complete symbol set contains a total of 150 unique symbols, shown in Table 1. Some of the maatras are collected more than once at different possible relative positions w.r.t. the underlying consonant. Including these, the total number of symbols is 166. These are all assigned to Unicode characters.

[1] https://en.wikipedia.org/wiki/Telugu_language

Table 1: The 150 unique (out of the total 166) symbols in the Telugu Dataset



The dataset is downloaded from HP Labs India website [1]. This dataset contains approx. 270 samples of each of 166 Telugu "characters" written by native Telugu writers. The dataset contains wide variation of distinct characters because of different peoples' writing styles. The characters are made available for download as TIFF files. Some of these character images are very complex shaped and closely correlated with others.

Solution Statement

In this project, I am using Convolutional neural networks to recognize handwritten Telugu characters. The reason for this is that CNN are more effective at determining the important features in a given image and finding patterns within images by using filters to find specific pixel groupings that are important. So, by using CNN, I want to write an algorithm to recognize the characters in Telugu language with more accuracy and less training time.

Benchmark Model

I plan to compare the results with the previous works. I will compare the accuracy/mean-squared error to see which is more effective, as well as compare the speed of my method using CDNN and the previous used techniques. Shanthi et al. [4] use pixel densities over different zones of the image as features for an SVM classifier. Their system achieved a recognition rate of 82.04% on a handwritten Tamil character database. K. Mohana Lakshmi et al. [3] achieved a recognition rate of 87.5% on Telugu character dataset using HOG features and Bayesian classification.

Here I want to try to get accuracy greater than 90%. For this I am applying Convolutional Neural Networks which are more efficient than the methods used by the above authors.

Evaluation Metrics

The evaluation metric for the model will be transcription accuracy on the test images in the HP dataset of Telugu characters. Accuracy will be defined as correctly predicting handwritten Telugu character in the image. The model must correctly predict not only the number of digits present but also correctly identify each of those digits. As noted above regarding benchmarking against previous work results, I will evaluate it against some other models like handwritten Hindi/ Tamil/ Bengali character recognition in both accuracy and speed.

Project Design

Programming Language and Libraries

- **Python 2.**
- **matplotlib.** Open source plotting library for Python.
- **scikit-learn**. Open source machine learning library for Python.
- **Keras.** Open source neural network library written in Python. It is capable of running on top of either Tensorflow or Theano.
- **TensorFlow.** Open source software libraries for deep learning.

The algorithm mainly contains two parts:

- Training of image.
- Testing of image.

Training and Testing of image:

STEP 1-Read input image:

The basic step that is required to start the procedure is to select or consider an image for the classification. The image should be first checked of its format and made sure that it is saved in the library of the files of the software we are using for its execution along with function format. Image file saved in the software library: 'imagename.jpg'. The image is named and saved for its usage as an input for simulation.

STEP 2-Resize Image:

After selecting the input image, then resize the image i.e. [50 50].

STEP 3-Converting image into gray:

Image in jpeg form, is transformed into gray form for easy analysis. The extraction of the histogram values and zoning is done accurately in a gray image.

STEP 4-Converting image into binary.

STEP 5-Implement CNN Classifier.

STEP 6- Testing of CNN Classifier.

The first stage of the project will be to download and preprocess the HP dataset. The images will then be cropped and resized to uniform size suitable for use by the model. The network will consist of several convolution layers which may be followed by pooling or normalization layers, followed by fully connected layers, and finally terminate with softmax classifier. The classifiers will be trained to recognize the character in the image.

The final evaluation of the model will be determined by computing the transcription accuracy of the predictions made against the test set. A separate validation set will be used to evaluate changes to the model's architecture and hyperparameters.

References

[1] http://lipitk.sourceforge.net/datasets/teluguchardata.htm

[2] K. Vijay Kumar, R. Rajeshwara Rao, "Improvement in Efficiency of Recognition of Handwritten Telugu script", International Journal of Inventive Engineering and Sciences (IJIES), Vol.-2, No. 1, pp. 1-4, 2013.

[3] K. Mohana Lakshmi et al. "Hand Written Telugu Character Recognition Using Bayesian Classifier", International Journal of Engineering and Technology (IJET)

[4] Shanthi N and Duraiswami K, "A Novel SVM -based Handwritten Tamil character recognition system", Springer, Pattern Analysis & Applications, Vol-13, No. 2, 173-180,2010.