OCR of Devnagari Characters

Final Project Report for CS 6350

Yogesh Mishra Nishant Agarwal

1 Introduction

Optical character recognition (**OCR**) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text. It is widely used as a form of data entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation. It is a common method of digitizing printed texts so that it can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

OCR is a widely researched field and it spans over a lot of languages. English has been the widely researched language for OCR, but with time researchers have worked on different languages in order to digitize text. Each language has unique characteristics and vary widely from each other is many aspects. The task is very challenging, but the outcome of being able to digitize text and making it accessible to the world is a huge achievement.

We have taken up OCR of Devnagari Characters as our problem. It is interesting to us as even today there are millions of documents that are waiting to be digitized and some of them are in pretty bad conditions. We found the task very exciting and challenging as OCR would help the world get access to rare manuscripts and documents which would become easily and always accessible.

In this report we present a comparison of performance of different classifiers on 3 forms of feature extraction methods: Binarization, Histogram of Oriented Gradients and Zoning.

2 Work

Our dataset [1] contained 110 uniques characters, including complex characters and possible segmented characters.The first step towards our project was extracting features from our images. We used computer vision techniques using **OpenCV**. We began by converting images to grayscale and binarizing the pixel vales. We worked with two sets of image sizes, 32×32 and 64×64, to get a push forward. We got 1024 and 4096 features respectively.

Using the **KNN** and **SVM** code from our assignments we modified the code to perform Multilable clasification. We used a One-vs-All approach. The accuracies received on the extracted features were 3% and 8% respectively for $k = 1$ and $C = 15$. We read upon on research papers about the ways of extracting features for handwritten characters and found that simple binarization is never used as features and that we would have to perform complex computer vision techniques on the binarized images to extract features. We selected two methods: **HoG**(Histogram of Oriented Gradients) and **Zoning**. HoG gives returns the pixel density values of the image based on the number of bins and block sizes. Zoning is the concept of finding the number of intersections, horizontal and vertical lines, distinct markers such as curves etc.

We started with small number of features,324 and 85 features respectively for each feature extraction method. At this point we modified our **Aggressive Perceptron** code to perform multilabel classification. We performed experiments using our modified KNN, SVM and Agressive Perceptron. We used scikit library to also run test using Naive Bayes.

We faced a lot of problem during experiments as we were not getting very good accuracies. We used cross validation for Perceptron and SVM to find the best parameters. We used $[0.0001, 0.001, 0.01, 0.1, 1.0, 2.0]$ as our rates, $[0.1, 1.0, 0.25, 3.5, 4.0, 0.05, 0.2, 0.4, 0.3, 0.5]$ was used for margins and $[10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$ was used as values for C.

3 Results

4 Conclusion

5 Future Work

References

[1] HP Labs. Hp labs india indic handwriting datasets. `http://programmingcomputervision.com/downloads/ProgrammingComputerVision_CCdraft.pdf`.