OCR using Machine Learning Techniques and Hidden Markov Models

1 Project Proposal

1.1 Project Idea

We would like to tackle the problem of Optical Character Recognition (OCR) using Machine Learning Techniques we have learned in class. The goal of OCR is to turn images into text. OCR can be analyzed using the noisy channel model, with the source being the human mind, and the channel being the handwriting technique. We want to determine the most likely letter given the handwriting.

For this project, we intend to design a Hidden Markov Model. The hidden states of the HMM will be the letters the human is thinking of, and the observed output is the actual handwriting represented as a pixel vector. This HMM can be graphically represented as a Bayes net. The goal is to find the sequence of hidden states (i.e. letters) that maximize the joint probability of those characters and the pixel vectors. Various OCR techniques (neural networks, naive bayes, logistic regression) can be applied to determine the probabilities that a pixel vector is a certain letter. We also intend to use the Viterbi algorithm to determine the most likely sequence of letters.

1.2 Data Set

We will be using the OCR dataset found at

`http://ai.stanford.edu/~btaskar/ocr/`

1.3 Software We Will Write

1. Adapt an algorithm for performing OCR that can be used on the dataset.
2. Write code for building HMMs and performing probability analysis on them.
3. Framework for integrating the above mentioned parts.

1.4 Teammates and Work Division

Jiang Lingzhang (lingzhaj) and Jonathan Yee (jyee1).

1. Read up on papers. (Both)
2. Work on concept and math related to HMM and its application to the dataset. (Both)
3. Look at data-set and write code for processing it. (Lingzhang)
4. Implement and analyze different OCR algorithms. (Both)
5. Compare and analyze results. (Both).

1.5 Midterm Milestone

We hope to implement OCR algorithms and run it on the dataset with satisfactory accuracy

References

[1] Ivan Dervisevic (2006) *Machine Learning Methods for Optical Character Recognition.* `http://perun.pmf.uns.ac.rs/radovanovic/dmsem/completed/2006/OCR.pdf`

[2] Paolo Frasconi, Giovanna Soda, Alessandro Vullo (2001) *Text Categorization for Multi-page Documents: A Hybrid Naive Bayes HMM Approach.* `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.3940&rep=rep1&type=pdf`