**EECS 348 Introduction to Artificial Intelligence**

**Members: He Zhao (hzg972), Xiaolin Li (xlo365), Zijiang Yang (zyz293)**

The features that we have tried are length (discrete), sum of the curvature (continuous), height-width ratio (discrete), speed (discrete) and box area(discrete). We also tested the possible combinations of these features to see if they can improve the accuracy of the classification. Our final decision is to use 4 features:

1. length(discrete)
2. sum of curvature (continuous)
3. height-width ratio (discrete)
4. box area (discrete)

While determining the appropriate thresholds for the three discrete features, we firstly decided the number of the thresholds that we needed for each of the features. For instance, for the height-width ratio, we need two thresholds to distinguish the text and the drawing because the height-width ratio for the text is always around one while that for the drawing is not. Then we used the featureTest() function to gather the values of the features for all the training data. We plotted the histogram on each of the features and set the value of the threshold which could best distinguishes the classes. (If there are two peaks separated by a deep valley, the feature value at the bottom of the valley is selected as the threshold.)

After we finished the development of the code, we did two tests:

==============================================================

**Trained and tested on the whole set of data**

**1. HMM basic (Where only one feature -- length is considered)**

Confusion matrix: {'text': {'text': 367, 'drawing': 374}, 'drawing': {'text': 351, 'drawing': 923}}

Percent\_correct: {'text': 0.49527665317139, 'drawing': 0.7244897959183674}

**2. HMMbest (Where length (discrete), Sum of curvature (continuous), Height-width ratios (discrete) and box area (discrete) are considered)**

Confusion matrix: {'text': {'text': 416, 'drawing': 325}, 'drawing': {'text': 285, 'drawing': 989}}

Percent\_correct: {'text': 0.5614035087719298, 'drawing': 0.7762951334379906}

**Trained with 50 files and tested with 20 files (15 overlapped)**

1. HMMbasic

Confusion matrix: {'text': {'text': 128, 'drawing': 177}, 'drawing': {'text': 33, 'drawing': 325}}

Percent\_correct: {'text': 0.419672131147541, 'drawing': 0.9078212290502793}

2. HMMbest

Confusion matrix: {'text': {'text': 163, 'drawing': 142}, 'drawing': {'text': 42, 'drawing': 316}}

Percent\_correct: {'text': 0.5344262295081967, 'drawing': 0.88268156424581}

From the comparison of case 1 where all the date were used as both training and testing data, the accuracies of the classification for both the text and the drawing classes have been improved.

For case 2, the accuracy of the text classification has been significantly improved (+0.12) without losing too much accuracy of the drawing class (-0.02).

From the results, we found that the set of additional features was a good quantification of the text features, but not so good for the drawing class. In addition, when we studied the impact of individual features, we found that the box area could improve the text classification accuracy and the height-width ratio can help the drawing classification. From the combined result, the power of the box area feature overwhelmed that of the height-width ratio feature using the independence assumption given by the getEmissionProb() function. We thought that, by modifying the getEmissionProb() to take the dependence of the features into account, the performance could be improved.