

Words and Pictures HW 3

(Piyush Kansal and Vinay K)

Part 1

For this part, we did following:

- segregated the data from all 4 bag categories into training set and test set by storing them in **trainingSet.mat** and **testSet.mat**
- calculated SIFT feature for each image from all 4 bag categories and stored them in **allSiftVectors.mat**
- using k-means clustering, clustered the features in **imageFeatureMap.mat**

Part 2

For this part, we did following:

- used stop-words to block words of no use. We got the file containing stop-words from a NLP project conducted by UC Berkeley, and it is quite exhaustive
- used Hash tables for faster processing and comparison of strings to classify training and query images and their associated descriptions
- create a lexicon of unique words from all 4 bag categories in **allLexicon.mat**
- then, using this unique word lexicon, calculated the histogram of word count from all the description files in both training and test data set and stored that in **allLexiconVector.mat**
- then, using the allLexiconVector, calculated the most frequent 1000 words and stored them in **mostFrequentLexicon.mat**

Part 3

For this part, we did following:

- Trained the Naive Bayes classifier for each feature and each category using the .mat files created so far
- calculated the probability of features for each bag category

Part 4

For this part, we did following:

- Classified the test images using our training data by calculating the probability for a image to be in a category and assigned it a category depending on the highest probability

- created the confusion matrix to depict the required behaviour and following is the o/p of the program

Code Organization

We have organized the code as follows:

- hw3.m
 - main program file
- initStopWords.m
 - initialized stop word hash table for faster lookups
- removeStopWords.m
 - remove stop words from the description files
- strip_punctuation.m
 - downloaded from internet to strip punctuation (REF: http://www.beamreach.org/soft/Xbat-win-Matlab/XBAT_PRE_R3/Core/Misc/strip_punctuation.m)
- computeSift.m
 - computes SIFT features
- initLexiconVector.m
 - initialized the lexicon vector
- findMaxCategory.m
 - finds the best category based on the maximum probability

Execution and Output of the Program

```
>> hw3('C:\Users\pkansal\Documents\SBU\Sem 3\Words and  
Pictures\hw3\newbags.tar\bags\');
```

ans =

347	125	18	9
55	393	23	28
11	29	453	6
22	158	24	295

Data Files

We have shared the location of all of the above mentioned data files and is located at:

https://docs.google.com/folder/d/0B9_MOqSCUfGINVIsWGFYdXhydM/edit