Randomly Sampled Codebooks vs. KMeans Codebooks

Duy Vu

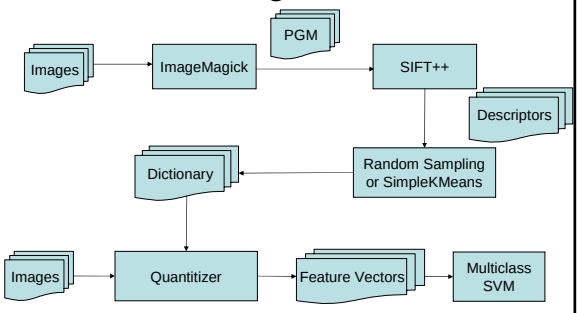
Goals

- Comparison on the performance of a randomly sampled dictionary and KMeans dictionary
  - Nowak, Jurie and Triggs. Sampling Strategies for Bag-of-Features Image Classificaiton, ECCV 2006
- How fast could I develop a system for image categorization from available tools?

Tools

- **SIFT++**:
  - http://vision.ucla.edu/~vedaldi/code/siftpp/siftpp.html
  - A lightweight C++ implementation of SIFT feature extractor.
  - ImageMagick converts images to the PGM format.
- **Weka** Machine Learning Package:
  - KMeans.
  - Multiclass SVM (SMO).
  - Other clustering and classification algorithms are also available.

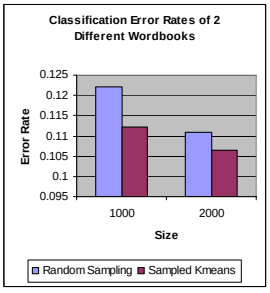
Diagram



Dataset

- Caltech-04
  - Background
  - Cars (rear)
  - Houses
  - Motorbikes (side)
  - Camels

Results



Size	% Error Reduction
1000	1.12
2000	0.57

Confusion matrices

Randomly Sampled Dictionary

dictionarySize = 1000

1320.0	0.0	49.0	0.0	1.0
0.0	120.0	2.0	4.0	0.0
195.0	2.0	**759.0**	18.0	26.0
5.0	7.0	27.0	767.0	20.0
30.0	0.0	48.0	15.0	263.0

dictionarySize = 2000

1290.0	0.0	76.0	0.0	4.0
0.0	118.0	3.0	5.0	0.0
139.0	5.0	**806.0**	18.0	32.0
7.0	4.0	24.0	773.0	18.0
15.0	1.0	44.0	18.0	278.0

Confusion matrices

KMeans Dictionary

dictionarySize = 1000

1308.0	0.0	61.0	0.0	1.0
0.0	121.0	1.0	4.0	0.0
167.0	1.0	**796.0**	10.0	26.0
3.0	3.0	17.0	783.0	20.0
26.0	0.0	50.0	18.0	262.0

dictionarySize = 2000

1277.0	0.0	91.0	0.0	2.0
0.0	121.0	2.0	3.0	0.0
125.0	3.0	**823.0**	17.0	32.0
8.0	4.0	22.0	780.0	12.0
13.0	0.0	40.0	18.0	285.0

Lessons

- The computational cost of KMeans is very high though I only use a sample of 60,000 SIFT descriptors for clustering.
- The time for quantization (creating feature vectors) is very long which is not usually mentioned in the papers.

Thank You!