11775 Homework 2 Report

Tianyu Xu - tianyux@andrew.cmu.edu

Stage 1

In this part, I followed the TA's instructions and did several experiments with several extracted visual features. The key steps are as follows:

1. I chose to use SIFT spatial Bag-of-Words as features to train the model. I noticed that all the feature files are in the same folder, so instead of using `select_files.pl`, I directly converted all the SIFT feature files into CSV files that `train_svm.py` could read in.

2. I used the file lists from homework 1 so I did not generate training / testing file lists again.

3. I noticed that some video clips in training / testing lists did not have corresponding SIFT feature file, and some SIFT feature files did not have the same dimension as others (32768). So I add modified train_svm.py so it first checks whether the SIFT file exists, then truncates the feature vector and just keep the first `feat_dim` columns.

4. The default SVM kernel of sklearn is rbf kernel, so I used different kernels instead.

5. Based on `run.med.sh`, I modified the script so that it could automatically read in all the converted feature CSV files, train a SVM model, use this model to classify the testing data, and use given mAP calculator to generate mAPs.

The final experiment results are as follows:

Kernel	Feature Dim	P001 mAP	P002 mAP	P003 mAP
RBF	300	0.175214	0.195617	0.209402
Chi-square	300	0.305362	0.677042	0.415476
Chi-square	32700	**0.518924**	0.67604	**0.568932**
Additive Chi-square	32700	0.496416	**0.736707**	0.497704

All the scripts and testing results can be found at: `/home/tianyux/hw2_stage1`

Stage 2

In this part, I extracted the SIFT features from raw videos, performed clustering, BoW generation, pooling, and got final presentations of videos. Then I trained an event classification model based on these presentations, and use the model to classify the test videos.

1. Similar to Stage 1, instead of using `select_files.pl`, I directly converted all the raw videos into keyframes. First I used FFmpeg to select the first 30 seconds of the videos and downsize them to 160×120 and 15 frames per second.

2. Then I used FFmpeg to select all the keyframes of videos:
   ```
   ffmpeg -y -i downsample/VIDEO.mp4 -vf
   select="eq(pict_type\,PICT_TYPE_I)" -vsync 2 -f image2
   keyframes/VIDEO_%03d.jpg
   ```
   On average, I got 30~40 keyframes for each video.

3. I wrote my own SIFT feature extractor and use it to extract all the SIFT features of each video. The extractor is called `sift_extract.cpp` and I followed TA's instructions to compile it.

4. I wrote a feature combiner to combine all the features in different keyframes of the same video.

5. I selected 20% of SIFT vectors and concatenate them into one file, and used this file to do k-means clustering. Then I used these clusters to generate BoW representation of all the videos.

6. I trained a SVM classifier, used it to get the rank of testing videos, and computed mAP based on output rankings. The results are shown as follows.

Kernel	Feature Dim	P001 mAP	P002 mAP	P003 mAP
RBF	400	0.262573	0.515589	0.310268
Additive Chi-square		0.371563	0.612141	**0.31175**
Chi-square		**0.387718**	**0.647182**	0.307319

All the scripts and testing results can be found at: `/home/tianyux/hw2_stage2`