Homework 1 Audio-based Multimedia Event Detection

Task

The task of this homework is to perform multimedia event detection (MED) with audio features. It's due on **Feb 2, 2015**.

What to Submit

You are required to build MED detectors with two feature types: MFCCs and ASR transcriptions. You need to report your separate mAP values on MFCCs and ASR transcriptions respectively. You will get extra credits if you do combination to fuse these two types of features together.

In addition to the mAP values, you are also required to submit a summary, describing the steps you take to get your numbers. The path to your experiments should be specified in the summary.

Data

This part is the same as data selection described in Homework 2.

The dataset contains 3 events:

P001 assembling_shelter
P002 batting_in_run
P003 making_cake

In the following file,

/data/MM2/MED/labelsets/old/csv/MED10TRN_20101215_JudgementMD.csv

you find the ID of all the videos and if a video is a positive example of event 001, 002 or 003, or a negative example (null video). Take event 001 as an example. You should parse the file to generate an index file with the ID of the first 10 positive videos for this event. Meanwhile, you are required to choose the first 100 negative examples from MED10TRN_20101215_JudgementMD.csv and incorporate their ID into your own index file of training videos. Similarly, to generate your own index file of testing videos, parse the ID of all the remaining positive videos for this event and the ID for the second 100 negative examples from MED10TRN_20101215_JudgementMD.csv. The same procedure can be used to generate the index file for event 002 and 003. You should also generate the label file for the

training and testing setsin this step (1 indicates a positive example whereas 0 indicates a negative example).

MFCC Features

1. Here are the tools used in this task. **These tools are already installed on the rocks cluster**. You can find their locations by looking at the example script.

ffmpeg : extract the audio tracks (.wav) from the video files

```
> ffmpeg -y -i HVC1110.mp4 -f wav HVC1110.wav
```

speech_tools : extract one channel from the audio file

```
> ch_wave HVC1110.wav -c 0 -o HVC1110.C0.wav
```

openSMILE : extract MFCC features

```
> SMILExtract -C config/MFCC12_0_D_A.conf -I HVC1110.C0.wav.wav -O
HVC1110.mfcc.csv
```

2. An example setup

To familiarize you with the overall pipeline, we create an example setup: /data/MM1/11-775/ymiao/HW1/example_setup

This example does MFCC extration, k-means clustering, bag-of-words representation. However,

* it does NOT mean that you have to follow this setup. **We encourage you to create your own setup**
* using this setup without any changes/improvement **will get you low grades**. There are many other things you could explore for better MAPs.

ASR Transcriptions

The ASR transcriptions are in /data/MM1/11-775/asr. Each video has a CTM file. For example, the ASR corresponding to HVC3715.mp4 is /data/MM1/11-775/asr/HVC3715.ctm

Each line in the CTM files has 6 fields. From left to right, these fields denote:

#1 Video ID
#2 Channel ID. Always 1
#3 Starting time of the word, in terms of seconds
#4 Duration of the word, in terms of seconds
#5 Word
#6 Confidence score. Always 1

For example, “HVC3715 1 2.02 0.23 do 1” means that the word “do” is spoken from 2.02s to (2.02+0.23)s in the video HVC3715.mp4

If you use the bag-of-word representation, you may only need to care about the #5 field, that is, the words.