CS2108

Assignment 2 Report

Natasha Koh Sze Sze
A0130894B

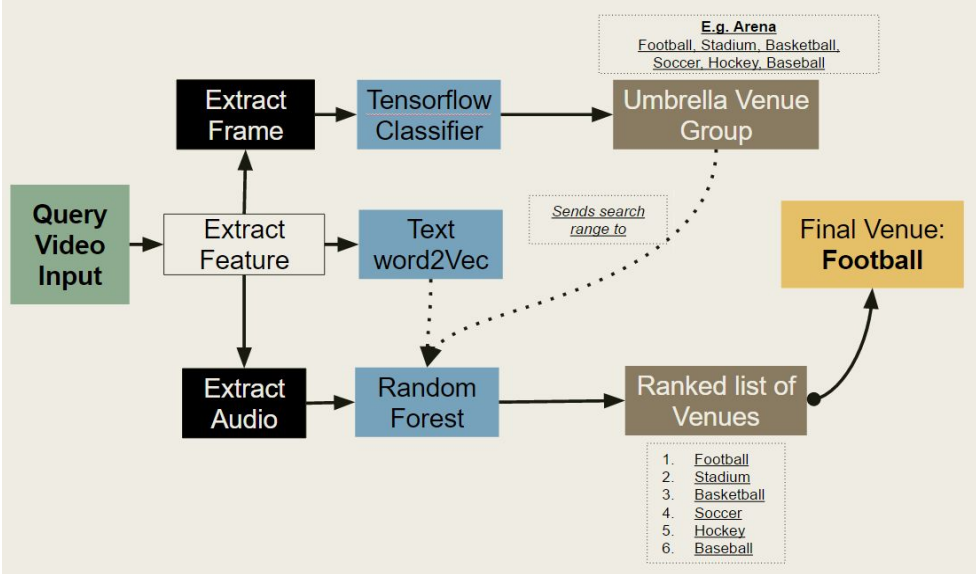
Chow Yuan Bin
A0108304U

Adrian Chan Ee Ray
A0122061B

Contents

Overall Design

1. Program Flow



`main.py` is the main class that generates the UI. Selected query videos are relayed to `querylogic.py` so that its frames, audio and text can be extracted and sent to the various classifiers for venue category prediction.

Audio, Visual, Textual Features

2.1 Acoustic

For the individual acoustic features, we first extract the audio from the provided video clips, followed by extracting the individual acoustic features from the audio clip. These individual features are the MFCC, the Mel-spectrogram, the Zero-crossing rate, and the RMSE of the audio. As not all audio clips have the same length, we were confronted with feature vectors between different audio clips having a different number of columns, which we opted to rectify by padding the feature vector of the shorter vector with zeroes.

2.1.1 Weighted Fusion of Acoustic Features

We adopted a weighting scheme for the acoustic features used in our retrieval engine, namely the MFCC, Mel-spectrogram (Magnitude Spectrum), Zero-crossing and Energy audio features. Below, we present our findings for the optimal weights for our retrieval engine.

Beginning with the initial values of MFCC = 0.5, Mel-spectrogram = 1.0, Zero-crossing = 0.75, and Energy = 1.0 due to our results when using the individual features on their own, we found an average F1 score of 0.0925. We then attempted

several different variations of the initial weights, where each weight could either be halved, doubled, or remain the same.

By taking these different combinations and determining the average F1 score of these combinations, we found the best result to be achieved when **MFCC = 0.5, Mel-spectrogram = 2.0, Zero-crossing = 0.75, and Energy = 1.0**, with an average F1 score of 0.1. Attempts to also increase the weight of Energy to 2.0 as well, despite its better individual accuracy, proved to return not as accurate a result, with a final F1 score of 0.0975. This may have been due to the better results from Energy, which stems from music venues (elaborated in section 3.2), not affecting the other venues as much, resulting in a lower than expected result.

2.2 Visual

For the visual features, we extracted keyframes from each video clip. A total of 5 frames were extracted per frame which are evenly distributed throughout the video to ensure an even distribution of frame results. These keyframes are then used by Tensorflow to determine where the video might be located in. Our visual classifier predicts a group of venue categories which a particular video might be classified as. This result will then allow the audio to focus its search to the specific groups of venues to speed up and improve accuracy or results.

2.3 Text

Text accompanying the video, from both the dataset and the input videos, are first combined into a single text file. Then, Word2Vec is run on this text file to generate a model that is suitable for the contents of the file, and a series of vectors for each video is generated. Finally, these vectors are separated back into dataset and input vector files. This allows us to directly compare the vectors of the dataset and the input files, since they were generated using the same model.

2.4 Combination and Fusion of Audio, Visual and Text

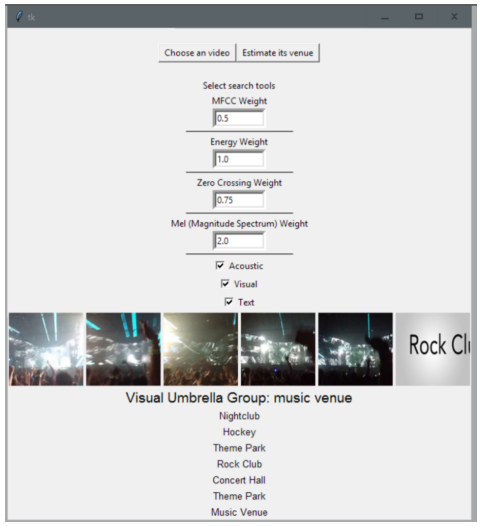
We implemented a voting system help determines the best venue prediction of an input video. First, the visual classifier returns a group of venue categories that best matches a video. Thereafter, we will choose the best venue from this narrowed pool of venue categories by looking at the list of venues returned by the audio and textual classifiers. This list is ranked by the confidence that a video belongs to that venue category. The venue that is has a higher vote will be returned as the final result. Additionally, our retrieval engine provides the flexibility to allow independent selection of these features should a user not want to search for all of these features at once.

2.5 Machine Learning Classifier

Our model of choice is the Random Forest Classifier with 500 tree estimators. The random forest fits our audio feature data into a number of decision tree classifiers and averages the classifications to improve predictive accuracy and to regularize over-fitting of our model out-of-sample. In following sections, we will analyse the default model given to us for this assignment and our Random Forest model, to show that the latter classifier enhances our retrieval engine's overall performance.

Testing

Below is a screenshot of an example video query input into our program.



Audio, Visual and Text features combined with acoustic weights

Analysis of Results

3.1 Approach

We first analysed the results of our audio classifier with all the venue categories to determine the best machine learning model to use. We found that using a Random Forest classifier was the best choice. We have tabulated our findings in 2 tables (Table 1.1 and Table 1.2) below, where we assess how the default model given to us and our Random Forest classifier perform on every venue category.

The Random Forest classifier gave us an overall higher F1 score of **8.5%** as compared to the default model at an F1 score of **4.25%**. The default model used a support vector machine with a radial basis function kernel of degree 3, while we used a Random Forest classifier with 500 tree estimators. We assume that the audio feature vectors used in the comparison of these 2 models have a hop length of 512 and used max-pooling to condense MFCC and mel-spectrogram audio feature vectors which had multiple rows.

The insights gained from this comparison allowed us to learn about how well our chosen Random Forest classifier model performs on each individual venue category and how different audio features in the environment affected our retrieval performance.

No.	Venues	Precision				Recall				F1			
		MFCC	MelSpect	Zero-Cross	Energy	MFCC	MelSpect	Zero-Cross	Energy	MFCC	MelSpect	Zero-Cross	Energy
1	City	0.04	0.09	0.12	0.11	0.03	0.07	0.13	0.07	0.04	0.08	0.12	0.08
2	Theme Park	0	0	0	0.25	0	0	0	0.03	0	0	0	0.06
3	Neighborhood	0.06	0	0.07	0	0.07	0	0.07	0	0.06	0	0.07	0
4	Other Outdoors	0.06	0.05	0.03	0	0.07	0.03	0.03	0	0.06	0.04	0.03	0
5	Park	0.04	0	0	0	0.03	0	0	0	0.04	0	0	0
6	Beach	0	0.12	0.05	0	0	0.07	0.03	0	0	0.09	0.04	0
7	Music Venue	0.04	0.07	0.05	0	0.07	0.03	0.03	0	0.05	0.04	0.04	0
8	Airport	0.06	0.11	0.21	0.1	0.07	0.07	0.17	0.07	0.06	0.08	0.19	0.08
9	University	0	0.07	0	0	0	0.3	0	0	0	0.11	0	0
10	Baseball	0.03	0.09	0	0	0.03	0.03	0	0	0.03	0.05	0	0
11	Home	0.09	0.14	0.2	0	0.1	0.1	0.33	0	0.09	0.12	0.25	0
12	Football	0.04	0.2	0.03	0	0.03	0.07	0.03	0	0.04	0.1	0.03	0
13	Stadium	0	0	0.07	0	0	0	0.1	0	0	0	0.08	0
14	Mall	0.03	0.25	0	0.06	0.03	0.2	0	0.03	0.03	0.22	0	0.04
15	Basketball	0.04	0.08	0	0	0.03	0.03	0	0	0.04	0.05	0	0
16	Art Museum	0.03	0.05	0	0.12	0.03	0.03	0	0.1	0.03	0.04	0	0.11
17	Concert Hall	0	0	0.04	0	0	0	0.03	0	0	0	0.04	0
18	Nightclub	0.1	0	0.06	0	0.07	0	0.07	0	0.08	0	0.06	0
19	Zoo	0	0.08	0.06	0.1	0	0.07	0.07	0.07	0	0.07	0.07	0.08
20	Entertainment	0.11	0	0.06	0	0.13	0	0.03	0	0.12	0	0.04	0
21	Travel	0.04	0.14	0.05	0	0.03	0.07	0.1	0	0.04	0.09	0.07	0
22	Plaza / Square	0.06	0.05	0	0	0.07	0.03	0	0	0.06	0.04	0	0
23	Hockey	0	0.27	0.1	0	0	0.13	0.1	0	0	0.18	0.1	0
24	Hotel	0	0.12	0	0	0	0.03	0	0	0	0.05	0	0
25	Rock Club	0.08	0.09	0	0.05	0.07	0.93	0	1	0.07	0.17	0	0.09
26	Landmark	0	0.04	0.04	0	0	0.03	0.03	0	0	0.04	0.04	0
27	Soccer	0	0	0.1	0	0	0	0.17	0	0	0	0.12	0
28	Historic Site	0.09	0.1	0.12	0.14	0.1	0.07	0.13	0.1	0.09	0.08	0.13	0.12
29	Resort	0.04	0.11	0	0.08	0.03	0.1	0	0.03	0.04	0.11	0	0.05
30	Other - Buildings	0	0	0	0.06	0	0	0	0.03	0	0	0	0.04
-	Each Average	0.04	0.08	0.05	0.05	0.04	0.08	0.06	0.05	0.04	0.06	0.05	0.02
-	Total Average	0.055				0.0575				0.0425			

Table 1.1: Default SVM RBF model classification scores for all venues

No.	Venues	Precision				Recall				F1			
		MFCC	MelSpect	Zero-Cross	Energy	MFCC	MelSpect	Zero-Cross	Energy	MFCC	MelSpect	Zero-Cross	Energy
1	City	0.04	0.05	0	0.09	0.03	0.07	0	0.1	0.04	0.06	0	0.1
2	Theme Park	0	0	0.12	0.12	0	0	0.07	0.17	0	0	0.09	0.14
3	Neighborhood	0.08	0.19	0.08	0	0.03	0.13	0.03	0	0.05	0.16	0.05	0
4	Other Outdoors	0	0.09	0	0.33	0	0.03	0	0.13	0	0.05	0	0.19
5	Park	0	0.08	0.14	0.08	0	0.03	0.1	0.03	0	0.05	0.12	0.05
6	Beach	0	0	0	0.03	0	0	0	0.03	0	0	0	0.03
7	Music Venue	0.09	0.12	0.15	0.35	0.13	0.13	0.07	0.43	0.11	0.13	0.09	0.39
8	Airport	0.04	0.03	0.16	0.06	0.03	0.03	0.17	0.1	0.04	0.03	0.16	0.08
9	University	0.04	0	0.11	0.07	0.03	0	0.07	0.07	0.04	0	0.08	0.07
10	Baseball	0.08	0	0.02	0.09	0.1	0	0.03	0.07	0.09	0	0.03	0.08
11	Home	0.19	0.21	0.27	0.14	0.37	0.2	0.4	0.23	0.25	0.21	0.32	0.18
12	Football	0.03	0	0	0	0.03	0	0	0	0.03	0	0	0
13	Stadium	0.02	0	0.14	0.07	0.03	0	0.4	0.1	0.03	0	0.21	0.09
14	Mall	0	0.07	0.06	0.04	0	0.13	0.03	0.03	0	0.1	0.04	0.04
15	Basketball	0.05	0.09	0.04	0.05	0.1	0.17	0.07	0.03	0.07	0.11	0.05	0.04
16	Art Museum	0.07	0.09	0.11	0.15	0.17	0.1	0.3	0.3	0.1	0.1	0.16	0.2
17	Concert Hall	0.06	0.04	0	0.07	0.03	0.03	0	0.07	0.04	0.03	0	0.07
18	Nightclub	0.08	0.09	0	0.24	0.03	0.1	0	0.3	0.05	0.1	0	0.26
19	Zoo	0.07	0.06	0	0.16	0.03	0.13	0	0.17	0.05	0.09	0	0.16
20	Entertainment	0	0	0	0	0	0	0	0	0	0	0	0
21	Travel	0.13	0.23	0.22	0.14	0.2	0.17	0.2	0.1	0.16	0.19	0.21	0.12
22	Plaza / Square	0.05	0.08	0.08	0	0.1	0.23	0.13	0	0.07	0.12	0.1	0
23	Hockey	0	0.23	0.07	0.12	0	0.1	0.07	0.03	0	0.14	0.07	0.05
24	Hotel	0	0.27	0	0.04	0	0.1	0	0.03	0	0.15	0	0.04
25	Rock Club	0	0.15	0	0.21	0	0.2	0	0.2	0	0.17	0	0.21
26	Landmark	0.08	0.12	0.08	0.06	0.2	0.13	0.17	0.07	0.12	0.13	0.11	0.06
27	Soccer	0.1	0.17	0.12	0.14	0.2	0.13	0.27	0.33	0.13	0.15	0.16	0.19
28	Historic Site	0.19	0.09	0.19	0.11	0.2	0.13	0.2	0.17	0.2	0.11	0.19	0.13
29	Resort	0	0.14	0	0	0	0.13	0	0	0	0.14	0	0
30	Other - Building	0.06	0.12	0	0.03	0.03	0.07	0	0.03	0.04	0.09	0	0.03
-	Each Average	0.05	0.09	0.07	0.1	0.07	0.09	0.08	0.11	0.06	0.09	0.09	0.1
-	Total Average	0.0775				0.0875				0.085			

Table 1.2: Random Forest model classification scores for all venues

3.2 Preliminary Analysis of Classifier Choice

With reference to Table 1.1 and 1.2 above, it can be seen that for the Random Forest model, it provides a better result as compared to the default model. We suppose that the reason for this is that our dataset contains numerous outliers or noise. Therefore, the default support vector machine classifier cannot separate data points as well as the Random Forest classifier as the dataset might not be linearly well-separated by its soft-margin hyperplane. Since we have a large enough dataset and given that a soft linear separating boundary does not fit our dataset well, we adopted the Random Forest classifier instead.

3.3 Analysis of Trends in Venue Category Data

From the above findings (Tables 1.1 and 1.2), we also note that for certain venues, such as “Music Venue”, “Nightclub”, “Rock Club”, using the Energy feature provided much better results as compared to the other features. This is likely due to these places having more easily identifiable energy levels, since all of these venues are expected to have some form of music playing in the background.

In addition, venues such as “Home”, “Travel”, “Soccer” and “Historic Site” perform better across the board. This is likely due to the more unique sound signatures to these venues, as no other venues have much overlap with these categories, except for “Soccer”. However, certain characteristics of soccer venues, such as the size and shape of soccer fields may present a different ambience as compared to other sports venues, such as the significantly smaller basketball court or the wider baseball fields.

In general, we note that Mel-spectrogram and Energy features perform better, while MFCC and Zero-crossing perform slightly worse for returning accurate venues.

3.4 Analysis of Audio Extraction Methods

Max pooling, Mean pooling or Concatenation

By observing how different methods of combining audio features affected our F1 score, we were also able to determine the best audio feature extraction method. These findings have been tabulated in Table 1.3 below, where we kept the parameter settings for each model the same and compared the retrieval scores for each type of audio extraction method “max pooling”, “mean pooling” or “concatenation”.

	Precision		Recall		F1	
	SVM (RBF)	RandomForest	SVM (RBF)	RandomForest	SVM (RBF)	RandomForest
Max Pooling	0.05	0.08	0.06	0.1	0.05	0.09
Mean Pooling	0.05	0.09	0.06	0.11	0.03	0.09
Concatenation	0	0.11	0.03	0.12	0	0.1

Table 1.3: Max-pooling, Mean-pooling and Concatenation performance

One key finding is that the concatenation of MFCC, Mel-spectrogram, zero-crossing and energy features performs better using the Random Forest classifier. It is interesting to note from our findings that it is a combination of audio features that produces a better representation of the audio sample. In other words, a combination of speech synthesis, sound pressure energy, short-time energy and detection of percussive sounds using the zero-crossing rate form a more accurate representation.

3.5 Analysis of Audio Classifier Parameters

We also observed that tuning our classifier's parameters and the parameters of our audio feature extractor changed our accuracy scores. We tabulated these findings as well in Table 1.4 below. With reference to Table 1.4, "hop length" represents the frame-to-frame time interval of our sampled audio waveform, while "n_estimators" refers to the number of tree estimators used in our Random Forest classifier.

	F1 Score											
n_estimators	5			100			500			1000		
hop length	4096	1024	512	4096	1024	512	4096	1024	512	4096	1024	512
MFCC	0.04	0.04	0.03	0.07	0.06	0.07	0.06	0.08	0.06	0.07	0.07	0.06
MelSpect	0.05	0.05	0.06	0.08	0.08	0.09	0.08	0.08	0.09	0.07	0.09	0.09
Zero-Cross	0.03	0.04	0.06	0.04	0.08	0.07	0.05	0.07	0.08	0.05	0.07	0.08
Energy	0.04	0.05	0.06	0.08	0.11	0.1	0.1	0.11	0.1	0.1	0.1	0.11

Table 1.4: Random Forest model parameter tuning

We noticed that the shorter the hop length/frame step size, the better the performance of our retrieval engine across all the various audio features. This could be due to the fact that with a smaller hop length, assuming we used the same sampling frequency, the frame rate at which we sample the audio waveform is larger. This would imply that the audio feature vectors extracted formed a more accurate representation of the audio source as more samples at every timeframe are analysed.

Also, using a larger number of tree estimators for our Random Forest classifier gives us a much better performance. However, at around 500+ trees, the increment in performance deteriorates greatly, hence 500 trees seems like the optimal number. We believe that this is largely because a larger number of trees allows us to obtain better out-of-sample estimates and it prevents the model from overfitting too much. Our audio sample dataset is quite large, which explains why our model allows for the use of a large number of trees and gives an improvement in performance.

3.6 Analysis of Visual Classifier

The training sets (total 15,000 frames) were sent to train over 6 hours with Tensorflow. However, we found that there is a low test accuracy (~30%). After proper investigation, we discovered that the provided training dataset had numerous invalid true positives.

Therefore, the training set has been painstakingly re-sorted (15,000 frames) to improve accuracy of results. Furthermore, instead of directly returning individual venues, we instead chose to return a collection of related venues, as accurately narrowing down the result to just one of these venues proved to be too inaccurate.

By grouping the data sets and re-sorting the results, the test accuracy has improved to 55% (Table 1.5) based on calculations of the recall and precision test. This grouping procedure allows the audio classifier to return a more accurate output.

No.	Venues	Precision	Recall	F1
1	Arena	0.6	0.7	0.6461538462
2	Music Venue	0.7	0.8	0.7466666667
3	Room	0.5	0.4	0.4444444444
4	Art Museum	0.4	0.42	0.4097560976
5	Zoo	0.81	0.8	0.8049689441
6	Mall	0.5	0.3	0.375
7	Outdoor Scenary	0.4	0.4	0.4
8	Public Spaces	0.4	0.4	0.4
9	Theme Parks	0.45	0.45	0.45
10	Airport	0.8	0.8	0.8
-	Total Average	0.556	0.547	0.5476989999

Table 1.5: Visual deep learning classifier performance

Results are particularly better for “Music venues”, “Arena”, “Zoo” and “Airport” as they are the more easily distinguishable and common visual features in the videos (airplanes, dark environment, open field, animals). It should also be noted that in the case of a wrong venue result, the second group venue is usually the correct venue.

Conclusion

We have assessed the performance of individual audio, visual and textual features and took an in depth look at how differing parameters of our audio features, audio extraction methods and machine learning classifiers affect performance at different levels.

We conclude that it is a combination of the various audio features that give a more accurate representation of the sound source and hence gives a better retrieval performance score. Furthermore, parameters that affect the extraction methods of audio signatures, the weights of different audio features and the classification of audio, visual and textual features allow us to better fine-tune our retrieval engine to perform better on a wider spectrum of venue categories.