


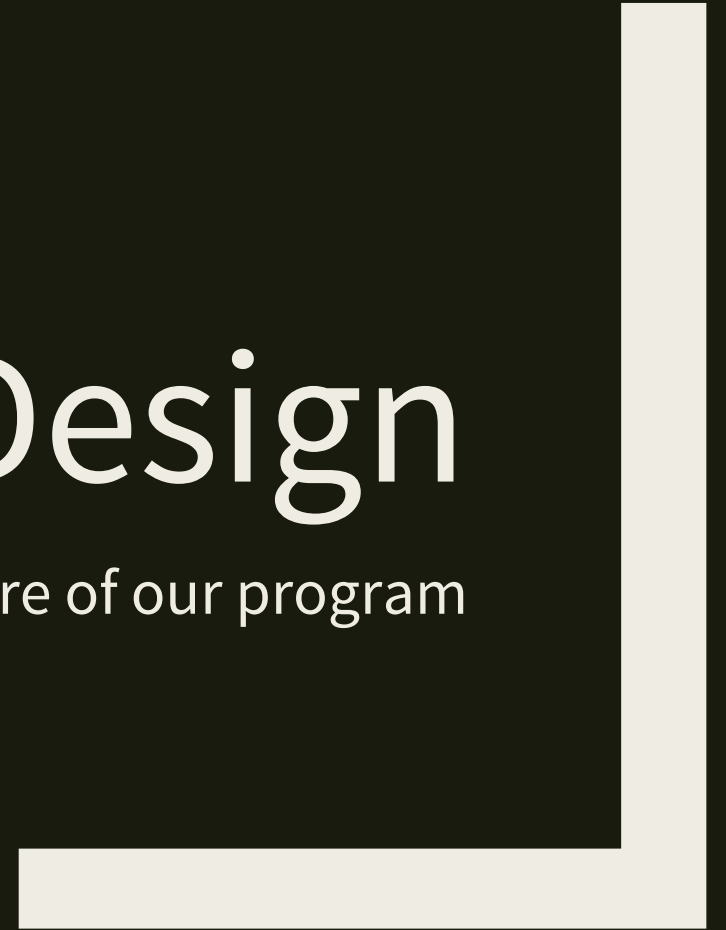
# CS2108

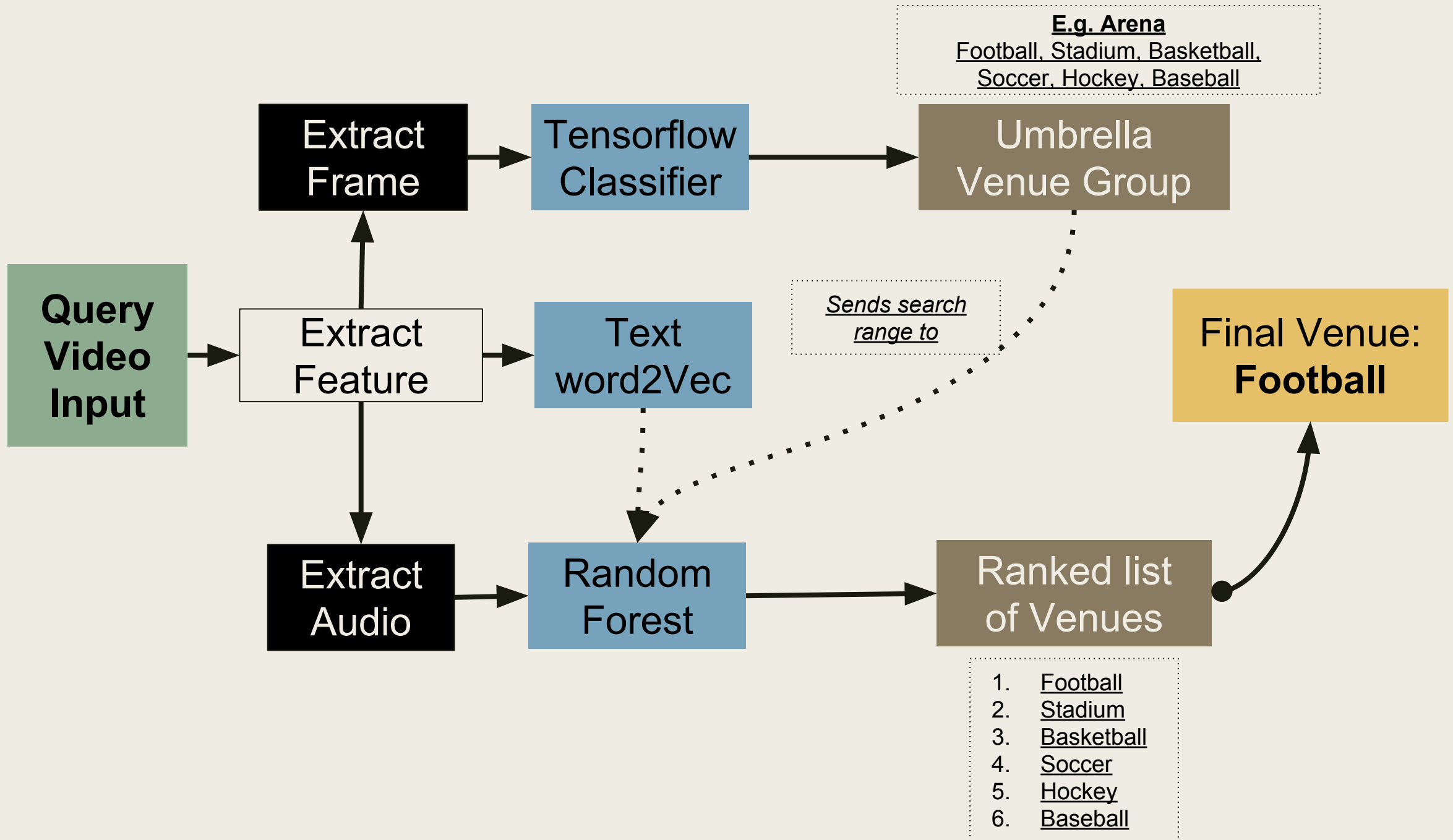
Assignment 2 Report



# Design

The architecture of our program





# Audio, Visual, Textual Features

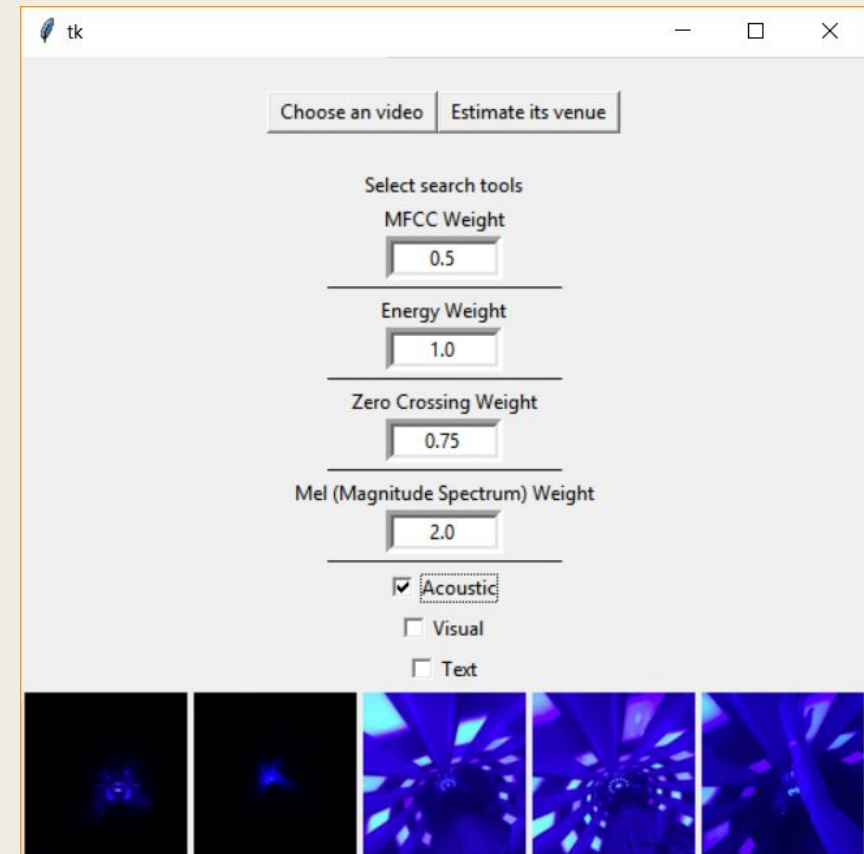
The specifics of our searching features



# Acoustics

-MFCC + Mel-spectrogram + Zero-crossing rate + RMSE

- Audios are extracted
- Vectors padded with zeros for videos with shorter lengths

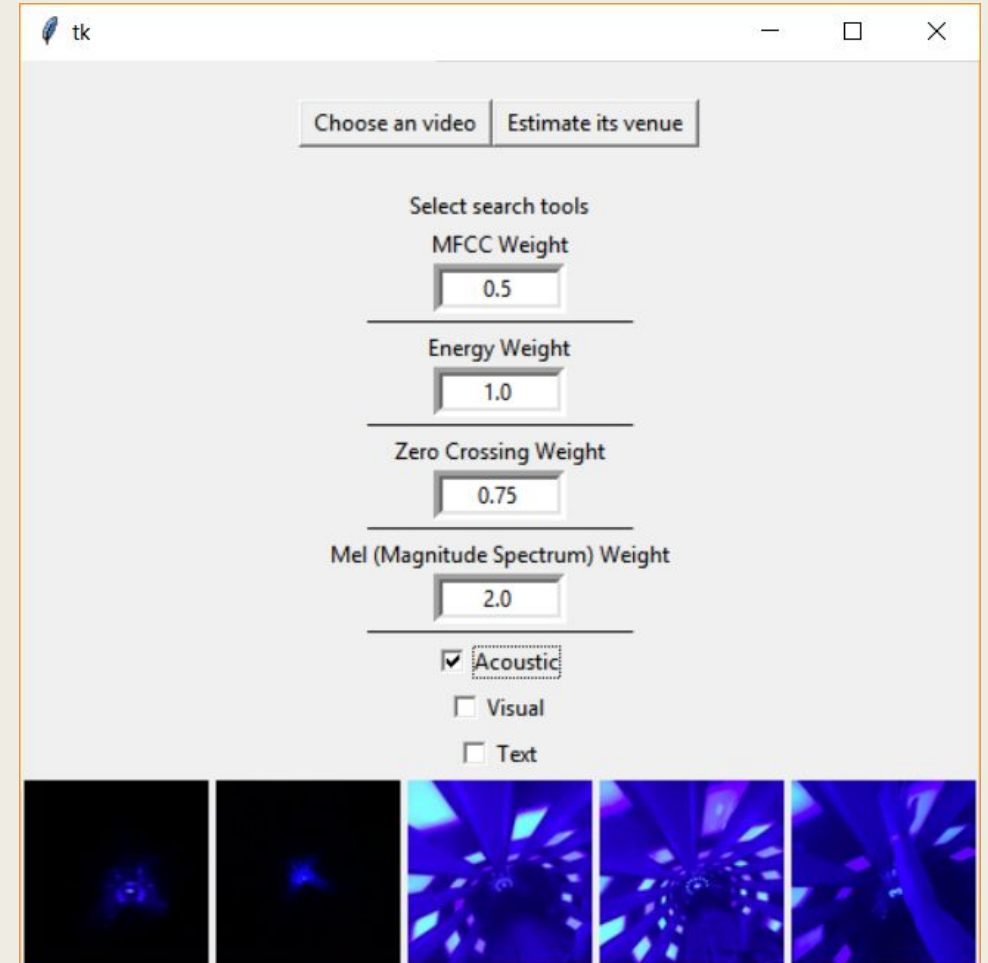


# Acoustics Weights

$a(\text{MFCC}) + b(\text{Energy}) + c(\text{ZeroCrossing}) + d(\text{Mel})$

= final venue result

- Average f1 = 0.1 (Energy = 1.0)
- Alt f1 = 0.0975 (Energy = 2.0)



tk

Choose an video | Estimate its venue

Select search tools

MFCC Weight

0.5

Energy Weight

1.0

Zero Crossing Weight

0.75

Mel (Magnitude Spectrum) Weight

2.0

☒ Acoustic

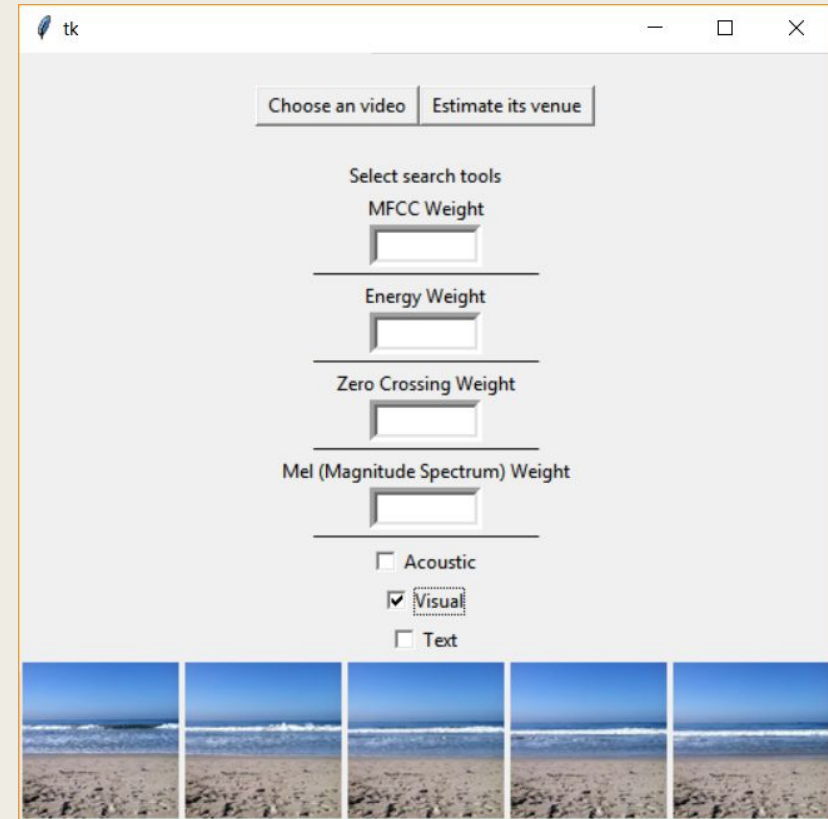
☐ Visual

☐ Text

Five small image thumbnails showing different venue interiors.

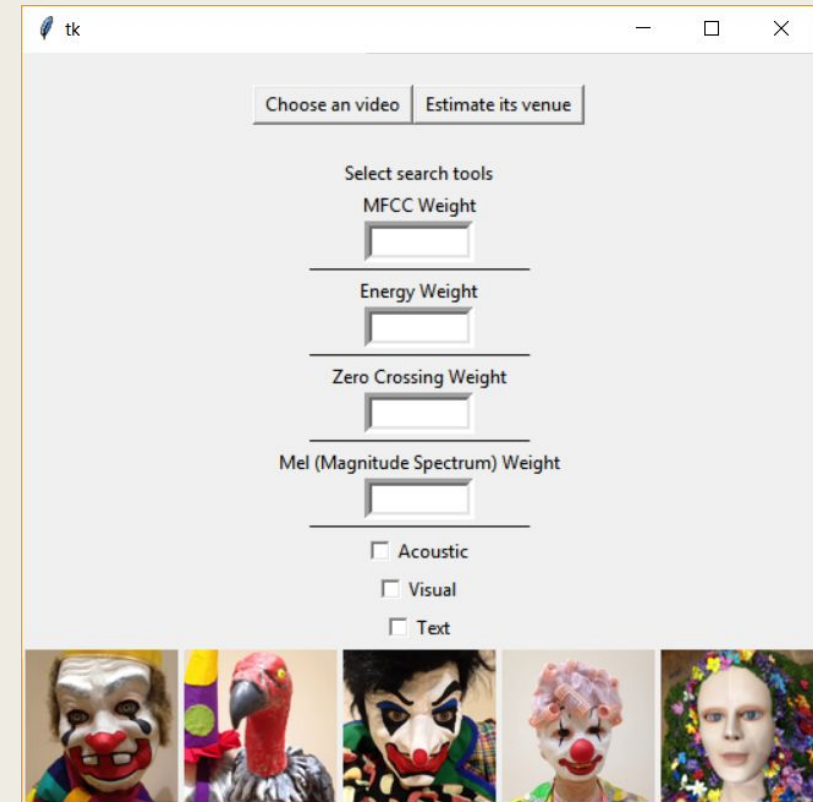
# Visual Classifier

- Tensorflow used
- 5 frames extracted per video for search
- Used to narrow search of venues by returning a result of “umbrella” venues
  - **Arena:** Stadium, Football, Soccer, Basketball, Hockey, Baseball
  - **Music Venues:** Music Venue, Nightclub, Theme Park, Rock Club, Concert Hall, Theme Park
- Results used to speed up venue classification



# Text Classifier

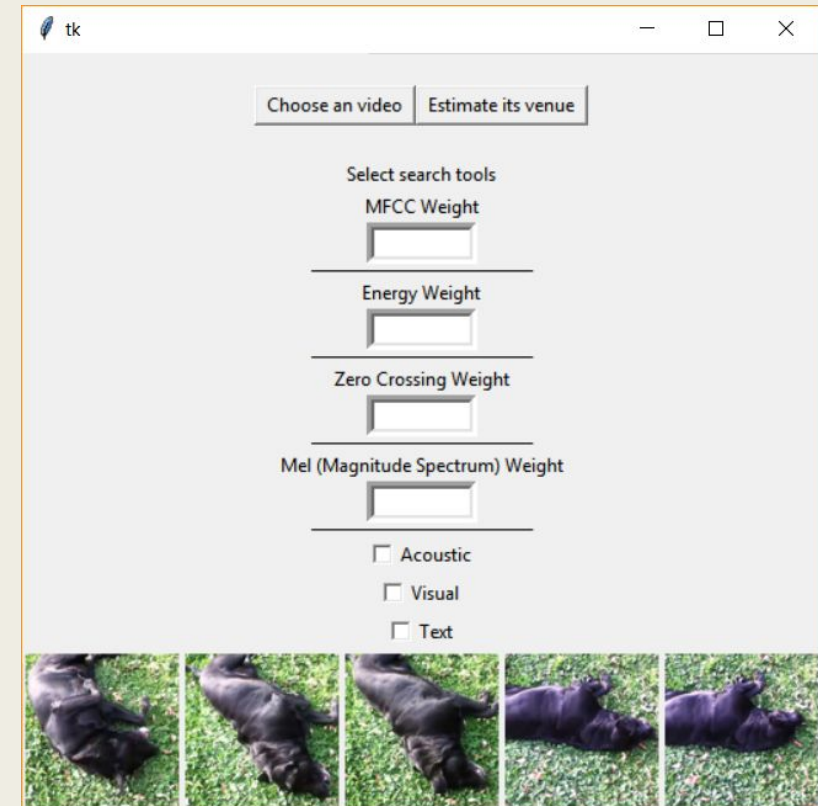
- Dataset and the input videos are combined into a single text file
- Generate model with Word2Vec, along with vectors for the videos
- Separated back into dataset and input





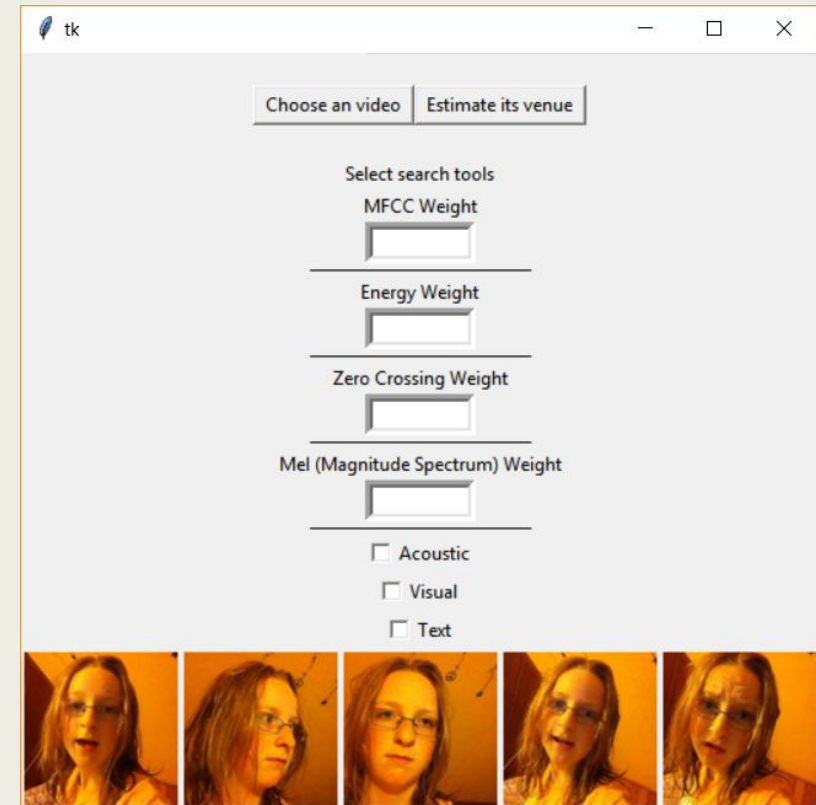
# Combination & Fusion of Classifiers

- Voting system determines best venue prediction
- Visual classifier narrows venue list
- Audio and text classifier ranked narrowed list by confidence
- Returns highest confidence vote



# Machine Learning Classifier

- Choice of classifier:  
**Random Forest**  
(500 trees estimator)
- Averages the classifications to
  - *Improve predictive accuracy*
  - *Regularize over-fitting of our model out-of-sample*



Analysis



# Default SVM RBF Model

Feature	Average F1
MFCC	0.04
MelSpect	0.06
Zero-Cross	0.05
Energy	0.02

You may refer to the report for a full page view

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			



# Random Forest Model

Feature	Average F1
MFCC	0.06
MelSpect	0.09
Zero-Cross	0.09
Energy	0.10

You may refer to the report for a full page view

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			

# Summary of venue analysis (audio)

- **Random Forest classifier performs better** and captures the patterns of audio features
  - Linear SVM might be deficient in separating data points linearly
- **“Music Venue”, “Nightclub”, “Rock Club”**, using the **Energy** feature provided much better results.
  - Distinct sound pressure energy levels from other venues
- **“Home”, “Travel”, “Soccer” and “Historic Site”** perform better across the board.
  - Unique sound signatures
- **Mel-spectrogram and Energy features perform better**, while **MFCC and Zero-crossing perform slightly worse**

# Analysis of Audio Extraction Methods

- **Concatenation** works best on Random Forest
- Result of combining audio features better than using an individual feature
  - *E.g. speech synthesis, sound pressure (energy), short-time energy, zero-crossing rates*

	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

# Audio Classifier Parameter Tuning

- **Shorter hop length**
  - Frame rate at sample audio is larger --> Better representation
- **Larger number of tree estimators for Random Forest classifier**
  - Prevents over-fitting and gives more reliable estimates

**Improved performance across the board.**

	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



# Visual Classifier Analysis

- **Training set (15,000 frames)**
- **Grouping procedure**
  - Initially (without grouping), low accuracy (~30%)
  - Current enhancement (with grouping), higher accuracy (~55%)
  -
- **Observations:**
  - “Music venues”, “Arena”, “Zoo” and Airport perform better.
  - Distinguishable and common visual features in vine videos

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

# Conclusion

- **Assessed performance of** audio, visual and textual features
- **Assessed performance of:**
  - *varying acoustic weights*
  - *extraction methods*
  - *audio-visual-text combination/fusion*
  - *machine learning classifiers*
- A **combination of features** (audio, visual and text) still gives the best result as it better represents all the dimensions of the human data.