Multimedia Feature Generation of Movie Trailers for Genre Prediction

John Fuini, Nathaniel Guy, and Yong Han Noel Kim
University of Washington, Seattle WA

***Abstract*— This paper presents a collection of techniques for the generation of visual and audio features based on film trailer data, and then shows a viable machine learning solution for building a classifier using these features which can identify the genre of a previously unseen film trailer. Our approach sought to capture, using computer vision and audio analysis techniques, the essential data within the trailer which is more characteristically informative as to the film's genre. These features included average intensity and color, number and length of distinct shots, amount of detail, amount of darkness, average volume, sudden changes in sound, and frequency distribution of sound. Using these features, along with genre metadata for each trailer, we trained both a Binary Decision Tree and a Support Vector Machine model to identify the genres. We applied extensive cross-validation and demonstrated classification accuracy greater than 65% for the top 10 most common genres, and greatrer than 90% for the top 5 most common genres. We also performed frequency analysis to study the modes on the feature set, and showed that the difference of success rate between one-mode classification and all-mode classification was less than 10%. We envision that our technique can be improved and applied to a variety of problems, and may be capable of competing with human accuracy on a similar problem.**

I. INTRODUCTION

Movie trailers are one of the most effective advertising tools for the film industry. They deliver relevant information such as background, cast, theme, plot and more, in a limited amount of time. As such, trailers can be considered a subset of a movie which contains its principal components. With this idea in mind, we developed an algorithm which classifies movie trailers by their genre. This is a familiar process to all movie-viewers: movie viewers are generally able to tell if a certain movie is a comedy film, action film, documentary, etc. within the first minute of watching a trailer based on myriad cinematic features within it. Viewers have developed this cognitive ability by watching countless movies of various genres over time, and have subconsciously learned to identify the cinematic features associated with certain genres. Our algorithm is an adaptation of this process using machine learning. This report describes our process of identifying cinematic features from a large set of trailers of known genre, training a machine learning algorithm to develop classification criteria based on these features and genre metadata, and testing the generated classification criteria on a set of trailers to assess the effectiveness of our approach.

Nathaniel Guy and Yong Han Noel Kim are Masters students in the University of Washington Department of Aeronautical and Astronautical Engineering, and can be reached at natguy@cs.washington.edu and kimber.noel@outlook.com, respectively. John Fuini is a PhD student in the University of Washington Department of Physics, and can be reached at fuini@uw.edu.

A. Sample Data Set

For the bulk of our classifier training and testing, we used trailer data from roughly 1,000 major motion pictures released within the last decade. These trailers were all high-resolution (the majority were 720p), with a typical framerate of 24 FPS. The trailers were downloaded from online sources using custom-built scripts, along with movie metadata (such as genre), which was scraped from webpage descriptions.

II. RELATED WORK

Zeeshan Rasheed et al., in their *On the Use of Computable Features for Film Classification*[1], developed an algorithm for film classification based on film previews. They limited themselves to visual features only, such as average shot length, color variance, motion content and lighting keys, and constrained their classification to four genres: comedy, action, drama and horror. In contrast, our work aims to create an algorithm that can classify a larger number of genres, using more features derived from both video and audio features, and with greater robustness than the technique in [1].

III. COMPONENT ARCHITECTURE

A. Feature Generation via Video Processing

We implemented a number of computer vision techniques to calculate features from video data. The majority of our video processing was utilized the OpenCV "cv2" module in the Python programming language [2].

1) Total Time/Number of Frames: OpenCV allows the processing of video data on a per-frame basis. Individual frames were counted, in order to get a total run-time of any given trailer in terms of frame count. OpenCV can provide the frames per second (FPS) for a given trailer as well, and this allowed us to calculate the run-time of trailers in seconds.

2) Average Intensity: Average grayscale intensity, across all pixels and all frames, was calculated using standard RGB weights for grayscale reduction:

$$Average\,intensity = 0.2989R + 0.5870G + 0.1140B$$

Other statistical metrics, such as intensity standard deviation, min, and max, were calculated as additional features. Note that certain regions of many trailers constituted a black letterbox, and could be excluded from this calculation after the determination that pixels at certain coordinate positions remain black throughout the entire duration of a trailer.

3) R, G and B Components: R, G and B components denote the proportion of red, green and blue colors for a given pixel. We calculated the average intensities along each of these color channels, across pixels and across all frames. (In order to minimize runtime memory requirements and simplify code structure, we elected to include the letter-box regions in this calculation for some of the statistical measures, such as color channel standard deviations and minimum/maximum intensities.)

4) Number of Shots: We detected a shot transition by examining and comparing color histograms of adjacent frames. When the chi-squared distance between two histograms exceeds a predetermined shot transition threshold, we determine that there was a shot transition between the two frames. One disadvantage to this method is that algorithm cannot differentiate between fast camera movement and complete change of shots, since with fast camera movement, the distribution of colors within the frame can vary as much as with an actual change of shot, especially with large moving objects. See Fig. 1 for an example of frames across a shot transition and their histograms.

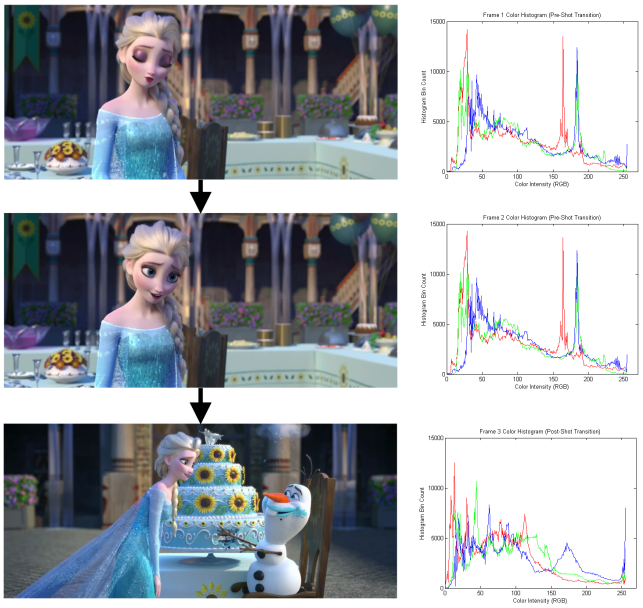

Fig. 1: Three example video frames are shown on the left, with their color histograms on the right. The first two frames are similar, and are part of the same shot; note their similar histograms. The third frame, only a fraction of a second later, has a starkly different histogram, and would thus be detected as a new shot.

5) Shot Length: Once the time-stamps of shot transitions were recognized, we were also able to calculate metrics based on the length of shots: mean, standard deviation, and minimum/maximum shot lengths. The mean shot length varied greatly on our sample set, from 0.2 seconds to 3 seconds.

6) Detail Score: We defined a measure of the amount of complex detail in a trailer through the use of the Canny edge detection filter [3]. By applying the Canny algorithm to a given frame, we can calculate a binary mask wherein all of the "edge pixels" in that frame have a value of 1. Then, by summing all of pixels in the frame, and summing all of the frames into the trailer, we can get a total "detail score" for that trailer. That score can then be scaled by the number of frames to normalize and find the average detail per frame. We calculated detail score features (including mean value, standard deviation, and minimum/maximum values) in this way. See Fig. 2 for an example of edge-detected images from which detail scores were calculated.

Fig. 2: Two example video frames are shown on the left, with their Canny edge-detected versions on the right. By counting the white (edge) pixels in the images on the right, a "detail score" can be calculated, as a rough metric of the detail complexity of the frame.

7) Dark Scenes: We defined the notion of a "dark scene" as a period of consecutive frames with very low average values (i.e., black transitions). Their lengths were recorded, from which mean length, standard deviation of length, and minimum/maximum lengths were calculated. The percentage of dark scene frames in each trailer was also calculated.

B. Feature Generation via Audio Processing

From the trailer videos, we isolated the audio component using a sampling frequency of 44.1 kHz. A series of analyses was performed on this audio data in order to extract features.

1) Volume: Mean: Mean volume for each trailer was calculated by averaging the amplitudes of sound waves over the entire duration of the trailer. One motivation behind extracting this feature was that a trailer saturated with loud noises would have larger value of mean volume than trailers with relatively calm sounds. Also, typically trailers with loud noises–explosions, jet noise, shouting, etc.–are associated with genres such as action, thriller, and adventure. On the other hand, trailers with calm audio, and even some quietness, may be associated with genres such as drama, history, and family.

Standard Deviation: all trailers were sourced from a single film site, but there was no guarantee that their audio was all equalized to the similar degree, especially because many were produced by different companies. Thus, a higher

mean volume could simply indicate a trailer that's louder in general due to different equalization characteristics, rather than an abundance of loud sound events. In order to get a sense of the variation of volume in each trailer, we calculated standard deviation of the sound wave amplitude over the entire trailer as well.

Minimum and Maximum: Minimum and maximum volume for each trailer were calculated based on the waveform data. On a scale from 0.0 to 1.0, most trailers had a minimum volume near 0.0, while some had marginally higher minimum values, such as 0.03. Maximum volume generally fell into the range of 0.1 to 0.5.

2) Sudden Rise/Fall of Volume: We defined the concept of a sudden rise and sudden fall of volume. Respectively, these represent an increase and a decrease of volume within a small time period, possessing a magnitude larger than the standard deviation of the volume across the whole trailer. Identifying the number of these events within a trailer's audio allowed us to study its audio dynamics, as we intuited that sudden increases of volume might be common during trailers that sought to startle viewers (such as those horror or action films).

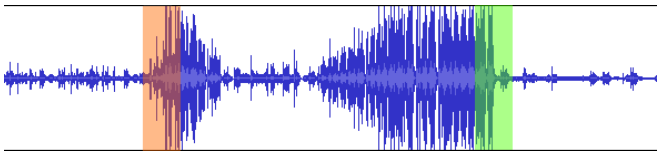

Fig. 3: An audio sample is shown, with a sudden rise and a sudden fall highlighted in orange and green, respectively. Note that the interior fall and rise are not captured as sudden changes, due to their more gradual change.

3) Percentage of Sound Corresponding to Different Octave Bands: For this feature, the waveform of each trailer was transformed to the frequency domain using the Fast Fourier transform (FFT). Its frequencies were divided into eleven bands in the audible range, commonly defined as octave bands (11Hz $\sim$ 22720Hz). The magnitudes of the frequency components in each band were summed together, and normalized so that the sum of magnitudes of all octave bands would be 1. The resulting binned magnitudes represented the composition of sounds of each trailer with respect to these eleven octave bands.

C. Use of Features in Machine Learning Algorithm

All extracted features from each trailer were compiled into a single comma-separated variable (CSV) spreadsheet. In addition to our generated features, the spreadsheet contained the genre labels for each trailer as well. Movies were not limited to one genre. For instance, there were movie trailers with multiple genre labels such as action-comedy, or mystery-horror-thriller. This spreadsheet was passed on to Matlab's fit binary classification decision tree function (*fitctree*) to build a decision tree. Only 80% of the trailers randomly selected from the full set of trailers were used for building the tree. This subset is known as a trainer set. The tree was then used to predict the entire range of trailers using Matlab's classification predict function (predict), and its success and failure rates were recorded. This process was repeated 40 times, each time with a new set of random trainer sets, for the purpose of cross-validation.

We experimented with the support vector machine (SVM) model as well. The classifier was created using MATLAB's SVM fit function (fitcsvm), used for classification, the results were cross-validated (crossval), and their accuracies were noted (kfoldLoss).

IV. RESULTS

The rates of success of classification by our algorithm for top ten most popular genres are shown in Fig. 4. Success rates were calculated using both binary decision tree, and support vector machine sub-algorithms. For instance, the Drama genre had 58% and 66% successful classification for binary decision tree and SVM respectively. It had the lowest success rate of top-ten-genres, but the rest genres registered consistently over 70% success rate. SVM had roughly 8% higher success rate for a given genre than binary decision tree.

In order to gain some insight to what features are important

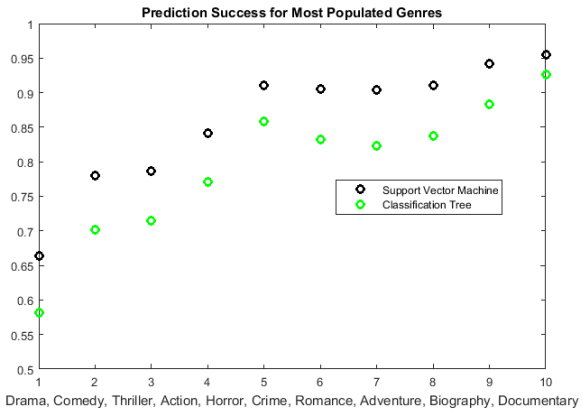

Fig. 4: Classifier success rate for ten most popular genres.

for the classification, we performed singular value decomposition on the set of movie trailers and their features. As can be seen in the plot of covariances of principal modes (fig. 6), one can see that there is no set of predominant modes that affect the decision, but its rather a complicated combination of all modes. The first four modes contain approximately 25% of energy in all modes. These four modes consist of features of differing degree. Figure **??** display the weights of each feature in four prominent modes. We could not identify any apparent patterns for any of these modes. We experimented with the concept of dimension reduction by trying the classification using one, four and all of the modes. The rate of success are plotted in fig. 7. Surprisingly, the difference of success rate between one-mode classification and all-mode classifications was less than 10%.

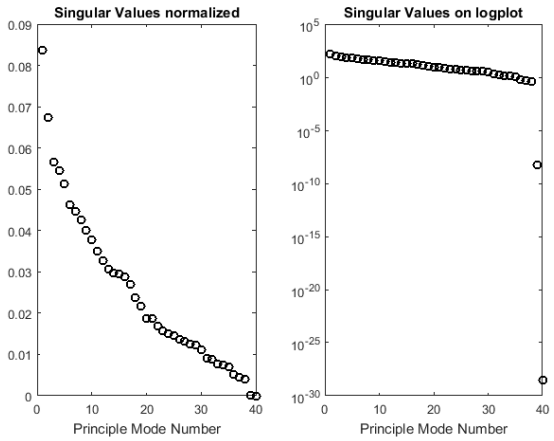

Fig. 5: Covariances of modes

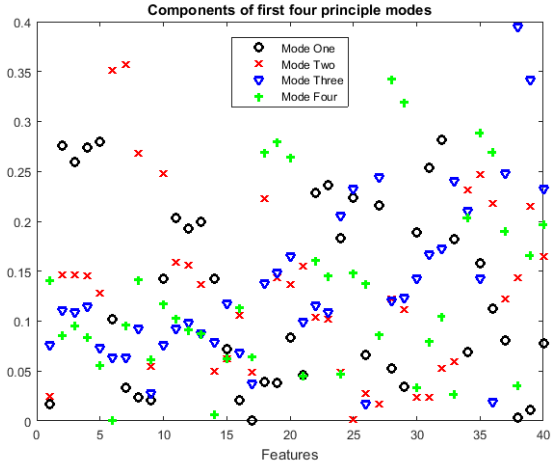

Fig. 6: Compositions of features for each modes

V. LESSONS LEARNT

A. Standardization of Features

One crucial mistake that was overlooked during the initial phase of our development was negligence to standardize features. The original features had different units, and different raw values, often differing by order of magnitudes. For example, features such as 'total run-time' or 'number of shots' had values in the range of hundreds, while features such as 'min. volume' or 'min. shot length' had values in the magnitude of one decimal points. Without standardization, variations of values present in features with large raw values seem much greater to Matlab's trainer than those present in features with small raw values. In addition, the mean value of many of the features were not zero. This resulted in erroneously result where those features with large values showed up as modes largely driving the decision tree. Once all the features were standardized, however, we achieved the results presented in the result section above. The principal modes and the composition of these modes became more complicated, but the algorithm treated features of different nature more fairly. The classification success rate was not affected much.

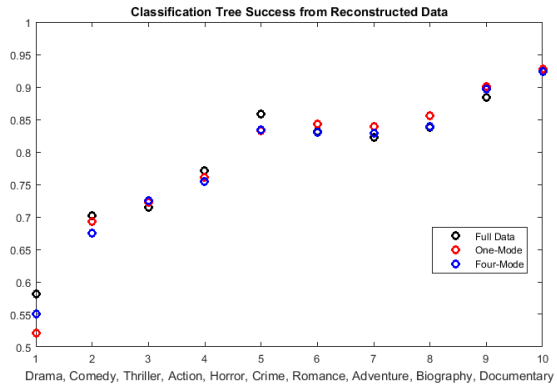

Fig. 7: Classifications made with one, two and four mode approximation compared to the original classification made with a complete set of modes

B. Unsuccessful Feature Extraction

One failed attempt to acquire features from the audio portion of movie trailers was to perform a Principal Component Analysis (PCA) using singular value decomposition. The goal of this process was to identify a series of principal modes and their components in each of the movie trailers. The hypothesized that the values of principal components could be used as features. We clipped 10-second portion of audio from each trailer to reduce the size of the data. Nonetheless, the matrix at which the singular value decomposition was to be performed had a size of 958-by-227150. This was computationally very expensive, with a runtime on the order of 10+ hours on modern personal computer. Furthermore, there was no guarantee that 10-second clipping would capture a signature sound of each trailer. (In a preliminary attempt, a 10-second clip was clipped at the timestamp of $t = \frac{total\ time}{2}$). Given these reasons, the attempt was deemed implausible in the sense of cost-benefit, and was abandoned.

VI. FUTURE WORK

An interesting topic of future work is comparing our results with the performance of human subjects. Since the definition of a genre is rather loose, it is often the case that there are many right answers for the genre of a given trailer. Thus, it is likely that human subjects guesses could differ from the genre labels in our metadata. We hypothesize that it may even be possible for machine learning to outperform humans at this task.

VII. CONCLUSION

This work demonstrated the construction of a movie genre classifier using features acquired from video and audio portions of movie trailers. A series of video analyses using a computer vision library generated numerous video features, such as detail scores, dark scenes, color and intensity profiles, etc. Temporal and frequency analysis of audio identified sharp increases and decreases of volume and frequency spectra. MATLAB classifiers trained using these features

classified movie genres with success rate ranging from 60% to 95%. Of two classifiers, the binary decision tree and support vector machine, the latter had approx. an 8% higher likelihood of successful classification overall.

It is worth noting that our machine learning algorithms pipeline has a minimum amount of required human interaction. The only step within that necessitates human input is providing a genre classification for the initial set of movie trailers used for training. The rest of the pipeline is almost entirely automated. While in this project we tested our algorithm using movie trailers, it can be adapted to any environment that requires computer vision-audio based machine learning. This can include surveillance, video tracking, and other applications.

VIII. ACKNOWLEDGMENTS

We would like to thank Dr. Nathan Kutz for his advice and feedback on our techniques used in this project. We'd also like to thank Kam Chancellor, because Nathan would want us to.

REFERENCES

[1] Rasheed Z., Sheikh Y., Shah M., *On the Use of Computable Features for Film Classification*, IEEE Transactions on Circuits and Systems for Video Technology, Vol.15 No.1, Jan. 2005.

[2] *OpenCV — OpenCV*. Itseez. Web. 09 Mar. 2016. http://opencv.org/.

[3] Canny, J., *A Computational Approach To Edge Detection*, IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679698, 1986.

APPENDICES

A. MATLAB codes

1) Frequency Analysis of Audio From Trailers:

```
%Feature extraction from audio portions of trailers.
clear all; close all; clc;

%Get list of files/ their names
folder_dir = 'C:\Users\Noel\git_repository\video-analysis-amath-582\video_processing\mp3s';
[mp3_path mp3_dir mp3_file] = dirwalk(folder_dir);
mp3_title{1} = regexprep(mp3_file{1}, '.\w*$', '');
mp3_title= mp3_title{1};
mp3_file= mp3_file{1};

%How many mp3?
total_n = length(mp3_title);

%Setup time, freq domain for FFT
sampling_freq=44100; %Sampling freq of 44100 fps, 10s.

band_binned_norm = zeros(total_n, 11);
%audible range of frequency is 20Hz to 20000Hz
%Human ears perceive pitch non-linearly. Use octave band

band_intervals = [11 22 44 88 177 355 710 1420 2840 5680 11360 22720];
ks_indices_for_cut = zeros(1,11);
for j=1: total_n
%Import audio
path_file = fullfile(folder_dir, mp3_file(j,1));
sound = audioread(char(path_file));
[mp3_length mp3_channel] = size(sound);

n= mp3_length;
L= mp3_length/sampling_freq;    %audio length in second
t2=linspace(0,L,n+1); t=t2(1:n);
k=(2*pi/L)*[0:n/2-1 -n/2:-1]; ks=fftshift(k);

%averaging L and R channel for brevity. No need if its already mono
if mp3_channel == 2
sound_avg = (sound(:,1) + sound(:,2))/2;
sound_avg = sound_avg(1:n,1)';
else
sound_avg = sound(1:n,1)';
end

%fft
sound_avg_t = fft(sound_avg);
sound_avg_t_s= fftshift(sound_avg_t);
sound_avg_t_s_abs = abs(sound_avg_t_s);

band_sum = zeros(11,1); %pre-populate
for jj=1:12
%find indices in ks corresponding to frequency bands (freq in Hz =
%ks(waveno. /2*pi
[idx idx] = min(abs(ks/(2*pi)-band_intervals(jj)));
ks_indices_for_cut(jj) = idx;
end
```

```
for jj=1:11
band_sum(jj,1) = sum(sound_avg_t_s_abs(1,ks_indices_for_cut(jj):ks_indices_for_cut(jj+1)));
end
band_total_sum = sum(band_sum);
band_sum_norm = band_sum./band_total_sum;
band_binned_norm(j,:) = band_sum_norm';
end
save('audio_freq_bin.mat', 'band_binned_norm','mp3_title','mp3_file','total_n')
```