CHARACTERIZATION OF MOVIES USING AUDIO VISUAL FEATURES

Gaganpreet Arora, Vinushree Chhajer, C V Jawahar

Center for Visual Information Technology,
International Institute of Information Technology, Hyderabad
Hyderabad - 500 032, INDIA

ABSTRACT

Characterization of videos with features and genres has many applications in multimedia mining. In the case of movies, it can help in recommendation systems, cinemetography etc. In this work, we observe and analyze patterns in the low level feature responses of videos. We show that the audio visual clues, which we compute, convey useful information about different aspects of movies such as genre, era, etc. We experiment on a database of 95 Oscar winning and nominated movies and discover many interesting patterns. Results are also presented for genre prediction of shots and movies.

Index Terms— Characterization of Videos, Classification, Genres of Movies

1. INTRODUCTION

Videos are becoming more and more popular for visual description of concepts, events, and dialogues. In general, videos can be divided into two distinct categories: videos with some content structure, and videos without any content structure [1]. The former class of videos, such as movies, broadcast news etc., conveys information in a structured manner. They convey a pre-defined content and meaning. The latter class includes videos like surveillance videos (monitoring for suspicious activity in airports, in wildlife census etc.). They typically have no scene change, therefore no content structure can be found in them. The content structure arises out of the association of video objects. They need to be characterized. By characterization, we mean, the process of representing a video (or sequence of frames) with a description which can aid a high-level reasoning at a later stage. In the case of movies, one would have liked to characterize them with high-level details related to characters, genre, plot, era etc. However, most of them can not be directly obtained from low-level description based on pixels and frames. In this paper, we attempt to characterize large number of movies, and discover some patterns by analyzing these features.

Characterization of movies can help in improving recommendation systems. However, popular recommendation systems are based on meta data of the movies, and not the audio-visual content [2]. Using a pre-classified database, user can be recommended movies based on movie patterns and his tastes. Another application is in the movie industry, for teaching cinematography students. The effect of color on the mood, music on the scene situation and post-processing of the audio and video on a human perception can be studied. These theories and practices of film aesthetics, can help a student in learning cinematic principles. Another important application is to mine patterns in movies for sociological studies. For example, one could mine pattern that over the years, the importance of roles played by females in movies have increased. However, applying mining techniques on movie data is difficult. Movies have rich semantics described using multimedia content; and mining them require meaningful characterization of the content.

Most of the previous related work has been done in categorizing videos as news, sports, sitcoms and commercials. Brezeale and Cook [3] have surveyed various literature on automatic video classification. The existing approaches generally focus on classification within two specific genres. For example, Moncrieff *et al.* [4] examine localized sound energy patterns, or events, that are associated with high level affect experienced with films. They experiment on 4 horror and 2 non-horror movies to establish a correlation between the sound energy event types and the horrific thematic content. Another similar type of classification is done by Jeho Nam *et al.* [5], where they classify violent scenes in TV and movies. Some interesting work has been done for movie characterization in [6, 7]. Authors investigate the problem of automatically labeling appearances of characters in TV or film material in [6]. The problem of aligning scripts to video/movies without subtitles is addressed in [7].

Most of the works discussed above, use limited set of shots or movies, while we look at a large collection of movies. Unlike these methods, we use low level features like shot length, color histograms and motion vectors for extracting useful information from a larger movie collection. Based on such feature responses from movies we discover and analyze some interesting patterns which can be used for characterization of movies.

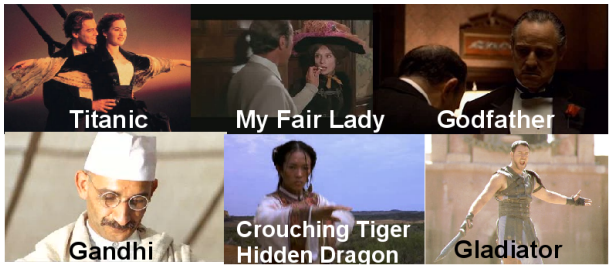

Fig. 1. Some Oscar wining and nominated movies in the dataset

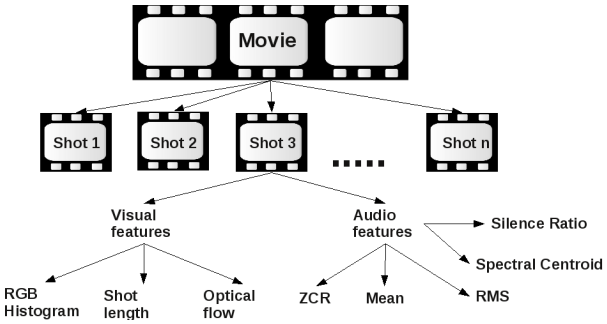

Fig. 2. Feature extraction process

2. DATASET AND FEATURES

2.1. Dataset

The dataset we work with includes 95 Oscar winning and nominated movies over the last 60 years ranging from 1950 to 2009. These movies form a good mixture and subset of the huge number of movies available. The genres, directors and other movie details of the dataset were annotated using the Internet Movie Database, IMDB, [8]. (Figure 1 shows examples of the movies).

2.2. Pre-Processing

To characterize movies, we identify and extract useful features on which database techniques can be applied. Our approach is divided into two sequence of steps, first shot detection which is followed by feature extraction. (See Figure 2)

Shot level features: A shot is a sequence of frames shot uninterruptedly by one camera. After dividing the movie into different shots, we extract visual and audio features on these shots. Shots are extracted by the method suggested by Zhang [9] *et al.* with α=5 and frame skipfactor of 2. Duration of shots is taken as one feature, and the visual and audio features are extracted on shots.

Visual features: Color histogram of a shot is one of the most common feature used. It can help know the lighting and time of the day a shot describes. $4 \times 4 \times 4$ bins of RGB space is used for computing color features. It is assumed that color within a shot will change minimally, hence a shot for color features can be represented using one keyframe. Optical flow or motion magnitude describe the transformation from one 2D image to another. Motion captures the amount of movement in a video and hence takes into account the temporal as well as spatial changes across the frames of a shot. Optical flow is computed using Pyramidal Lucas Kanade Optical Flow [10] method. The optical flow value for each shot is calculated as the mean of magnitude of motion vector over all the frames of the shot.

Audio Features: An advantage of audio features is that they require fewer computational resources than visual methods. We use five types of audio features, shown in Figure 2. Mean of absolute signal samples is the most common audio feature used. Root mean square, which is also called as standard deviation of signal, provides information about energy of the signal. Zero crossing rate (ZCR) is the number of times the signal amplitude changes sign in the current frame. Dialogs (or human speech) normally have a higher variability of the ZCR than in music. Silence ratio is the ratio of silence/noise to the actual signal. Spectral centroid is another measure to characterize an audio spectrum. It has a robust connection with the impression of "brightness" of a sound.

The number of detected shots for Oscar wining and nominated dataset is 73,626 with an average of 800.26 shots/movie. Average length of a shot is 9.37 seconds with 42,736 shots of less than 5 seconds duration, 26,998 between 5-30 seconds and 4,232 greater than 30 seconds.

3. PATTERNS ACROSS CATEGORIES

Over the last 50 years, movie making patterns have changed with the advancement of technology. Digital effects which were not possible a few years back are being used in all the movies now. Cameras can capture at faster frame rate and increasing use of safety measures allows for more stunts to be performed. Movies can be characterized in many different ways (i) based on period of picturization, (ii) based on mood of the film, (iii) based on location of picturization (iv) based on the schools of direction. (v) based on language. etc. Patterns exist which distinguish one type of movie or video in general from another. Some patterns that can be observed in regard to movie editing are,

- Movies of different genres - action, drama, comedy, horror etc. have different characteristics.
- Movies today have greater dynamics and action content than those picturized in the past which relied more on dramatization.
- Different directors have different cinematic principles they follow.

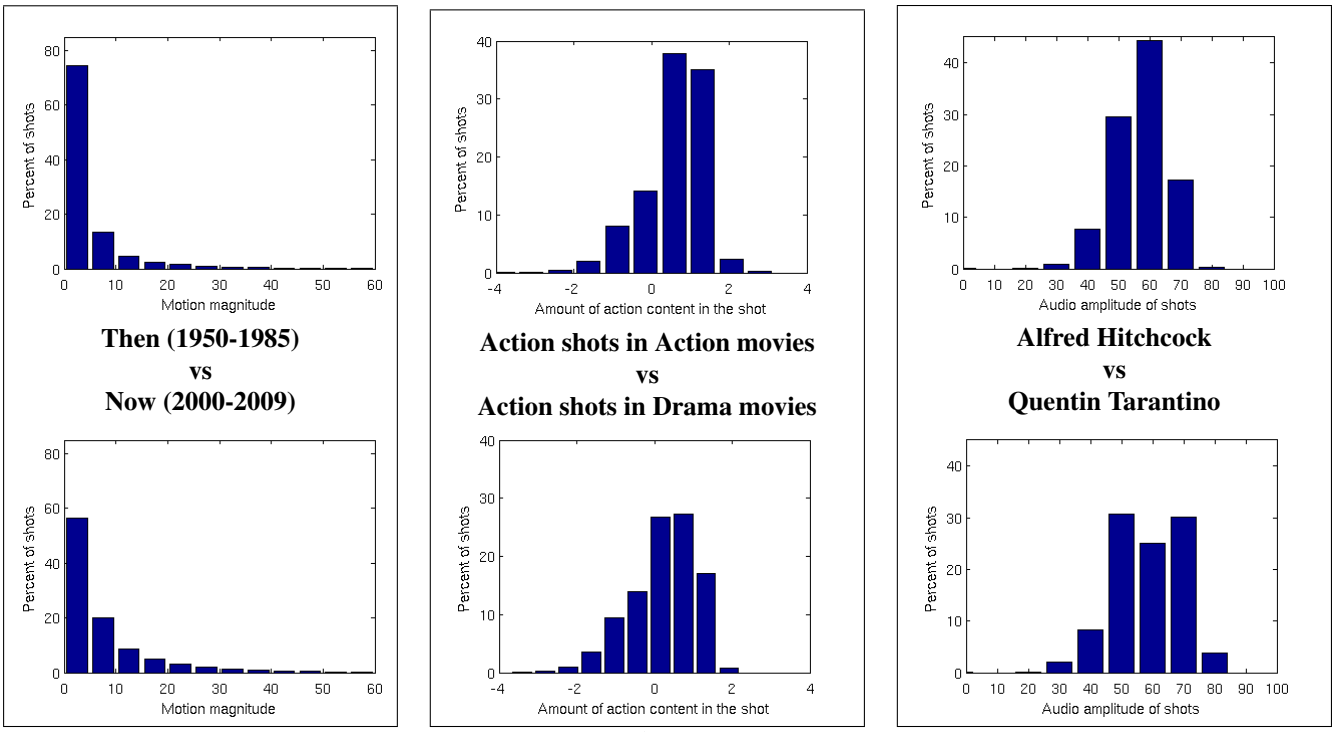

Fig. 3. Meaningful patterns in the movies - (a) compares "Then" and "Now" movies, (b) compares "Action" and "Drama" movies, and (c) compares "Hitchcock" and "Tarantino" movies.

Movies, Then vs Now: To validate our notions, we divided our dataset into two parts - old movies from 1950-1985 forming the "then" part, and newer movies from 2000-2009 forming the "now" part. Plots of motion and shot duration were plotted. The histogram plot for motion magnitude is shown in Figure 3 (a), with motion magnitude on the x axis, and the percentage of shots on the y axis. Around 70% of the shots of the "then" movies had negligible dynamics, whereas this has reduced to less than 50% in the "now" movies. Similarly, in shot duration, a higher percentage of long shots and a lower percentage of action content was observed in "then" movies.

Action vs Drama: Action movies have a large number of fighting sequences, explosions, crashes, accidents etc. Drama movies rely more on dialog and music (like crying, dance and tragic sequences) and less on dynamics. Based on this definition, our classifier gives a action:drama content for each shot of all the movies. High positive value denotes more action and high negative value less or no action, which is represented on the x axis. Two categories of movies - Action and Drama were created using their IMDB genre tag, and action percentage of shots was plotted. Figure 3 (b) shows Action movies have 70% shots with high action content, i.e. dynamism, while only 55% shots of drama movies have high action content. It was also observed that action movies have smaller shot duration than drama movies. Similary, a pattern of higher motion and dynamics in action movies was observed.

Alfred Hitchcock vs Quentin Tarantino: Another categorization of movies is according to the school of direction. We took 4 movies each of Alfred Hitchcock and Quentin Tarantino. Alfred Hitchcock was an English filmmaker and producer who pioneered many techniques in the suspense and psychological thriller genres. Quentin Tarantino is an American film director and producer whose films use nonlinear storylines and aestheticization of violence. We analyzed the patterns for each for both the directors and found that their style of direction is different, in spite of the fact that both of them are known for action and thriller movies. The histogram plot for audio magnitude is shown in Figure 3 (c), with audio amplitude on the x axis, and percentage of shots on the y axis. Around 45% of the shots of Hitchcock movies have higher audio amplitude and lower dynamics, whereas this is reduced to less than 30-35% in Tarantino movies. Similary, in shot duration, a higher percentage of long shots and a lower percentage of action content was observed in Hitchcock movies.

The above results clearly show that there exist a natural pattern which can be utilized in movie characterization. In section 4, we use these patterns as basis and provide genre classification for shots and movies.

Classifier	Accuracy before	Accuracy after
Action vs. Rest	77.91%	80.46%
Dialog vs. Rest	72.95%	79.8%
Drama vs. Rest	75%	77.25%
Dialog vs. Action	81.16%	84.12%

Table 1. Results for genre prediction of shots, before and after bootstrapping

4. GENRE CHARACTERIZATION AND PREDICTION

As an application of our approach, we perform experiments for genre characterization prediction for shots and movies.

Shot Genre Prediction: Shots are classified/tagged according to four classes taken by Sugano *et al.* in [11] and one extra category for music and songs. *(a) Action*, a fighting sequence, explosion, crashes etc. where lots of dynamics are observed in visual as well as audio stream. *(b) Drama*, a sequence where lots of tension and drama is seen (like people crying, running etc.) where a curiosity for an important event is being aroused with not much change in dynamics. *(c) Normal dialog*, a sequence where people are talking without any or little dynamics in visual and audio stream. Generally, the background music is low-monotonous or not present. *(d) Music/Songs*, a sequence where people are dancing/singing to celebrate. Conveying story and emotions through words and music with some defined rhythm can be termed as songs. *(e) Generic*, a shot that does not belong to any of the above mentioned categories.

About 1800 shots were randomly selected from our dataset and tagged manually in the above mentioned categories, with at least 300 in each class. Pair-wise and one vs rest classifiers were built for each of the above mentioned classes with all the tagged shots used for training and 60,000 untagged for testing. Bootstrapping (taking a percentage of highly negative shots as negative samples in the training dataset) was then performed to increase the quality of the results. The result accuracy (on average) before bootstrapping for 200 most positive predicted shots was 76%, and after bootstrapping it was 80%. Selected results are shown in Table 1.

Movie Genre Prediction: Each shot of the movie is tagged with a genre using the above classifiers (Action vs Rest, Action vs Dialog, Dialog vs Rest). The genre which occurs the most in a movie is then the predicted genre of the movie. A movie is labeled as Action, Drama, Action+Drama (nearly in equal proportion). Genre prediction of movies from shot genres gives an accuracy of 78.23%.

We show that genre of shots or movies can be characterized by using low level features analyzed in section 3. Thus, natural patterns exist and can be utilized in movie or shot characterization.

5. CONCLUSION AND DISCUSSION

We present an approach for movie characterization based on patterns present in the low-level feature responses. With the help of experimental study and analysis we support our argument. Our approach can be used along with other techniques to solve interesting problems like genre classification, movie success prediction, teaching cinematography, pattern mining.

6. REFERENCES

[1] Xingquan Zhu; Xindong Wu, "Sequential association mining for video summarization", Multimedia and Expo, ICME 2003.

[2] Meuth, R.J.; Robinette, P., Wunsch, D.C., "Computational intelligence meets the NetFlix prize", Neural Networks, IJCNN 2008.

[3] Brezeale, D.; Cook, D.J., "Automatic Video Classification: A Survey of the Literature", IEEE transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2008

[4] Moncrieff, S.; Venkatesh, S.; Dorai, C., "Horror film genre typing and scene labeling via audio analysis", ICME Multimedia and Expo, 2003.

[5] Nam, J.; Alghoniemy, M.; Tewfik, A.H., "Audio-visual content-based violent scene characterization", Image Processing, ICIP 1998.

[6] M. Everingham, J. Sivic, and Andrew Zisserman. "Hello! My name is... Buffy - Automatic Naming of Characters in TV Video", BMVC, 2006.

[7] Pramod Sankar, C. V. Jawahar, and Andrew Zisserman. "Subtitle-free Movie to Script Alignment", BMVC, 2009.

[8] Internet Movie Database, http://www.imdb.com

[9] Zhang, H., Kankanhalli, A., and Smoliar, S. W. "Automatic partitioning of full-motion video". Multimedia Syst., Jan. 1993,

[10] Lucas B D and Kanade T, An iterative image registration technique with an application to stereo vision. Proceedings of Imaging understanding workshop, 1981

[11] Sugano, M.; Isaksson, R.; Nakajima, Y.; Yanagihara, H., "Shot genre classification using compressed audio-visual features", Image Processing, ICIP 2003.