

Movie Genre Classification By Exploiting Audio-Visual Features Of Previews

Zeeshan Rasheed

Mubarak Shah

Computer Vision Lab, University of Central Florida, Orlando

E-mail: {zrasheed, shah}@cs.ucf.edu

Abstract

We present a method to classify movies on the basis of audio-visual cues present in the previews. A preview summarizes the main idea of a movie providing suitable amount of information to perform the genre classification. In our approach movies are initially classified into action and non-action by computing the visual disturbance feature and average shot length of every movie. Visual disturbance is defined as a measure of motion content in a clip. Next we use color, audio and cinematic principles for further classification into comedy, horror, drama/other and movies containing explosions and gunfire. This work is a step towards automatically building and updating video database, thus resulting in minimum human intervention. Other potential applications include browsing and retrieval of videos on the Internet (video-on-demand), video libraries, and rating of the movies.

1. Introduction

Movies constitute a large portion of the entertainment industry. Currently several websites host videos and provide users the facility to browse and watch movies online. Therefore the automatic classification of the movies on the basis of their content is an important task. For example movies containing violence or profanity must be put in a separate class as they are not suitable for children. Similarly, automatic recommendation of movies based on personal preferences will help a person to choose the movie of his interest. However, classifying a huge collection of video data without human intervention is not an easy task.

Movie directors often choose the most interesting and important events of the story to include in the movie previews to attract the viewers. A careful analysis of the movie previews can lead to an appropriate classification. For example [1] uses the *average shot length* and *shot activity* as features and classify movies into different genres like *action*, *romance/comedy* etc. In [2] authors identify violence in the trailers. We believe that low-level features such as color and music may be combined with the high-level domain knowledge to classify movie genre.

2. Our Approach

Directors often follow rules pertaining to the specific genre of a movie. Such rules are referred as *Film Grammar* or *Cinematic Principles* in the film literature. By following

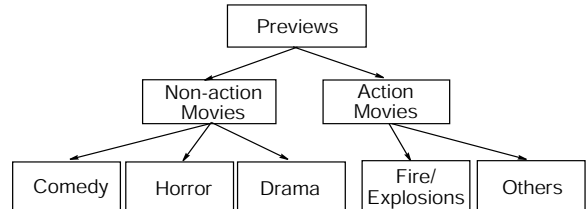


Figure 1. Flow chart showing the classes of movie genre

these principles, camera movements, sound effects and lighting can create mood and atmosphere, induce emotional reactions and convey information to the viewers. Although, different directors use these principles differently, movies of the same genre have a lot of features in common. For example, most of the action movies have similar shots and sound effects. Our aim is to analyze these audio-visual cues from the movie previews and make an educated guess about its genre.

We first classify movies into *action* and *non-action* classes by estimating the *visual disturbance* and *average shot length* using a very simple but robust technique. Visual disturbance is defined as the motion content of a video clip (Sec.2.1 and 2.2). We also make use of the color and audio information and combine that with the *Cinematic Principles* to classify movies. We make three subclasses; *comedy*, *horror* and *drama/other* under *non-action* group. Finally we classify *action* movies into *explosion/fire* category and *other-action* category (see Fig.1). This is done by first analyzing audio information. We find *events* by identifying the peaks in sound energy and test corresponding video frames for the occurrence of an explosion. Sec. 3 and 4 discuss the sub-classification. Sec.5 presents the experimental results and finally Sec.6 concludes our work.

2.1. Shot detection and Average shot length

We use a modified form of the algorithm reported in [3] for the detection of shot boundaries using HSV color histogram intersection. Let $D(i)$ represents the intersection of histograms H_i and H_{i-1} of frames i and $i-1$ respectively. That is:

$$D(i) = \sum_{j \in \text{allbins}} \min(H_i(j) - H_{i-1}(j)) \quad (1)$$

Then we define the shot change measure $S(i)$ as

$$S(i) = D(i) - D(i-1) \quad (2)$$

Shot boundaries are detected by setting a threshold on S . For each shot that we extract, the middle frame within the shot boundary is picked as a *key frame*. The average shot length is also computed by dividing the total number of frames by the total number of shots in the preview.

2.2. Visual Disturbance in the scenes

To find *visual disturbance* we use an approach based on the structural tensor computation introduced in [4]. The frames contained in a video clip can be thought of a volume obtained by combining all the frames in time. This volume can be decomposed into a set of two 2D temporal slices, $I(x, t)$ and $I(y, t)$, also called horizontal and vertical slices respectively. We evaluate the structure tensor of the slices as:

$$\Gamma = \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} = \begin{bmatrix} \sum_w H_x^2 & \sum_w H_x H_t \\ \sum_w H_x H_t & \sum_w H_t^2 \end{bmatrix} \quad (3)$$

where H_x and H_t are the partial derivatives of $I(x, t)$ along the spatial and temporal dimensions respectively, and w is the window of support (3x3 in our experiments). The direction of gray level change in w , θ , is expressed as:

$$R \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} R^T = \begin{bmatrix} \lambda_x & 0 \\ 0 & \lambda_t \end{bmatrix} \quad (4)$$

where λ_x and λ_y are the eigen values and R is the rotation matrix. The angle of orientation θ is computed as:

$$\theta = \frac{1}{2} \tan^{-1} \frac{2J_{xt}}{J_{xx} - J_{tt}} \quad (5)$$

When there is no motion in a shot, θ is constant for all pixels. In case of global motion the gray levels of all pixels in a row change in the same direction. This results in similar values of θ . However, in case of local motion, pixels that move independently will have different orientation. This can be used to identify each pixel in a column of a slice as a moving or a non-moving pixel.

We analyze the distribution of θ for each column of the horizontal slice by generating a nonlinear histogram of seven bins. In case of a static scene or a scene with global motion all pixels fall into one bin, whereas pixels with motion other than global motion have different values of θ and fall into different bins. The peak in the histogram is located and corresponding pixels are marked *static* while remaining ones are marked *moving*. The *visual disturbance* is the ratio of moving pixels to the total number of the pixels in a slice.

We use the average of disturbance of four equally separated horizontal slices to reduce the computational complexity. This measure is proportional to the amount of action occurring in a shot. In Fig.2 the density of *disturbance* is smaller for a *non-action* shot than that of an *action* shot.

2.3. Initial classification

We have observed that *action* movies have more local motion than a *drama* or a *horror* movie which results in

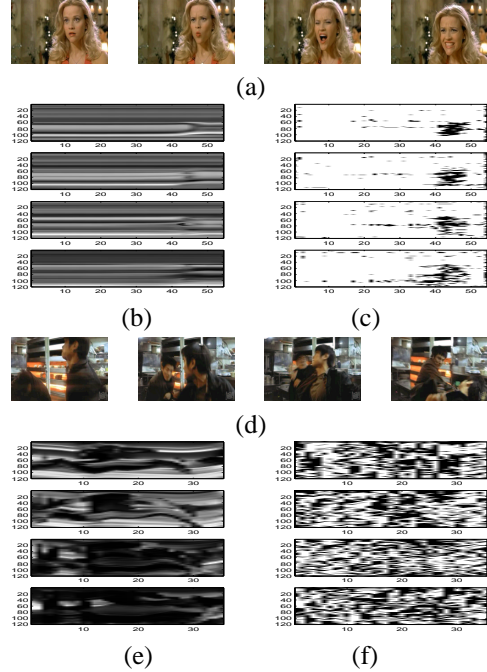


Figure 2. Plot of *Visual disturbance*, four frames of shots from (a) *Legally Blonde* and (d) *Kiss of the Dragon*. (b) and (e) are the horizontal slices for four fixed rows of corresponding shots. (c) and (f) show active pixels (black) in corresponding slices.

a larger *visual disturbance*. Also shots in *action* movies change rapidly than in other genres, *drama* and *comedy* for example. We plot the *visual disturbance* against average shot length and use a linear classifier to separate *action* movies from *non-action*. Similar conclusions have been drawn in [1].

3. Sub-classification of non-action movies

Light intensity in the scene is always controlled and changed in accordance with the scene. Movie directors balance the amount and direction of light while taking a shot to provide us a specific perception of the scene. Reynertson says “*the amount and distribution of light in relation to shadow and darkness and the relative tonal value of the scene is a primary visual means of setting mood.*” [6]p.107. Terms *low-key lighting* and *high-key lighting* are used in film literature to express the amount of light in the scene.

- **High-key lighting** The scene has an abundance of bright light with lesser contrast and the difference between the brightest light and the dimmest light is small. *High-key* scenes are usually happy or less dramatic. Many situation comedies also have high-key lighting.

- **Low-key lighting** The background and the part of the scene is generally predominantly dark with high contrast ratio. *Low-key* lighting being more dramatic are often used in *Film Noir* or *horror* films.

In *horror* movies shots are mostly *low-key*. On the other hand, *comedy* movies tend to have more *high-key* shots. We consider all key-frames of the preview in the gray scale space and compute the distribution of the gray level of the pixels. Our experiments show that:

(a) Comedy: Movies belonging to this category have a gray-scale mean near the center of the gray-scale axis, with a large standard deviation. This indicates a uniform distribution of light

(b) Horror: Movies of this type have a mean gray-scale value towards the dark end of the axis, and have low standard deviation. This is because of the frequent use of dark tones and dim lights by the director.

(c) Drama/other: Generally, these types of movies do not have any of the above distinguishing features.

For a preview i , we find the mean, μ , and standard deviation, σ , of gray scale distribution by using all key frames. We then define a quantity $\zeta_i(\mu, \sigma)$ which is the product of μ_i and σ_i , that is:

$$\zeta_i = \mu_i \cdot \sigma_i \quad (6)$$

Since horror movies have more *low-key* frames, both mean and standard deviation values are low, resulting in a small value of ζ . Comedy movies, on the other hand will return a high value of ζ because of high mean and high standard deviation. We define two thresholds, τ_c and τ_h , and assign a category to each movie i based on the following criterion.

$$L(i) = \begin{cases} Comedy & \zeta_i \geq \tau_c \\ Horror & \zeta_i \leq \tau_h \\ Drama/Other & \tau_h < \zeta_i < \tau_c \end{cases} \quad (7)$$

Figure 3 shows the distribution for three different sub categories of the movies.

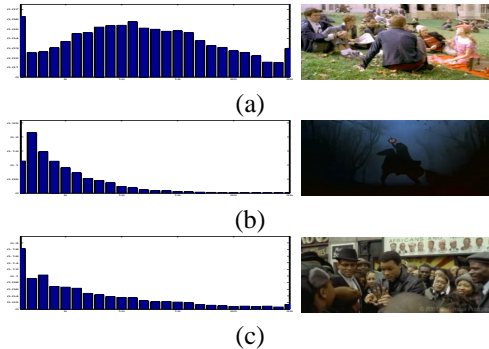


Figure 3. Average intensity histogram of key frames (a) Legally Blonde, a comedy movie (b) Sleepy Hollow, a horror movie and (c) Ali, an example of drama/other.

4. Sub classification within action movies

Action movies can be classified as *Martial art*, *War* or *Violent* (containing gunfire/explosions) etc. In this paper we further rate a movie on the amount of fire/explosions present

in its preview. We make use of both audio and video information to achieve this task.

4.1. Audio analysis

In Hollywood movies, music and nonliteral sounds are often used to provide additional energy to the scene. The audio is always correlated with the scene. For example, fighting, explosions, etc. are mostly accompanied with a sudden change in the audio level. Therefore the energy in the audio track is computed as:

$$E = \sum_{i \in interval} (A_i)^2 \quad (8)$$

where A_i is the audio sample indexed by time i . Interval was set to 50ms. We are interested in the instances where

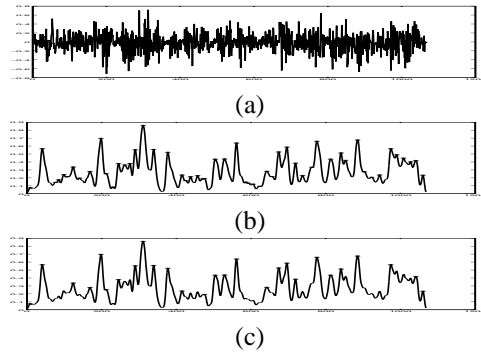


Figure 4. Audio processing: (a) the audio waveform of the movie The World Is Not Enough, (b) Energy plot of the audio: '*' indicates the peaks, (c) Good peaks are indicated by '*' after running the peakiness test.

the energy in audio changes abruptly, therefore, we perform a peakiness test on the energy plot. A peak is *good* if it is sharp and deep. Videos corresponding to the peaks above than a threshold, T_{peak} , are selected to process the corresponding video. See Fig. 4 for the plots of audio signal and its energy of *The World Is Not Enough*.

4.2. Fire/explosion detection

We analyze the video frames corresponding to *good* audio peaks for fire/explosion detection. In such cases there is a gradual change in the intensity of the images in the video from low to high. We compute the gray level histograms with 26 bins of the frames that are within the shot boundary of shot referred by peakiness test and plot the index of the bin with the maximum number of votes against time. In case of an explosion, the shot shows a gradual increase in the intensity. Therefore, the gray level of the pixels move from lower intensity to higher intensity values and the peak of the histogram moves from a lower index to a higher index. We exclude shots that show low stability in the plot since a camera flash which does not last for more than a few frames could be considered as an explosion. Fig.5 show the detection in two candidate shots.

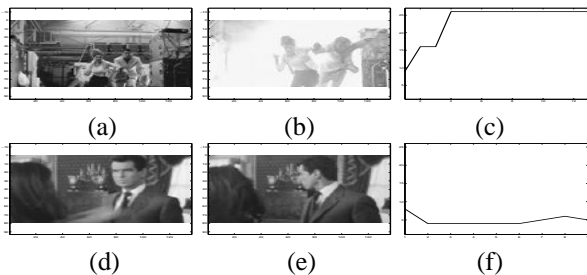


Figure 5. Detection of fire/explosion. (a)-(b) are frames of first shot and (d)-(e) are frames of second shot. (c) and (f) are the plots of index of histogram peak against time. First shot was identified as fire/explosion.

5. Experiments

We have experimented with previews of 19 Hollywood movies downloaded from Apple’s website [7]. Video tracks were analyzed at the frame rate of 24Hz and at the resolution of 120x68 whereas the audio was processed at 22KHz and with 16-bit precision.

Fig.6 shows the distribution of movies on the feature plane by plotting the *visual disturbance* against the *average shot length* and separating by a linear classifier. Movies with more action contents exhibit smaller average shot length. On the other hand *comedy/drama* movies have low action content and larger shot length.

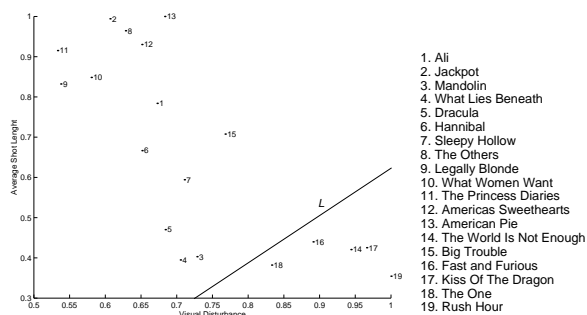


Figure 6. The distribution of Movies on the basis of Visual Disturbance and Average shot length.

Using the intensity distribution of key frames of *non-action* class, we label movies as *comedy*, *horror* and *drama/other*. See Fig.7. *Dracula*, *Sleepy Hollow*, *What Lies Beneath* and *The Others* were correctly classified as *horror* movies. Movies that are neither *comedy* nor *horror* including *Ali*, *Jackpot*, *Hannibal* and *What Women Want* were also labeled correctly. There is a misclassification of the movie *Mandolin* which was marked as a *comedy* although it is a *drama* according to its official website. The only cue used here is the intensity images of key frames. We expect that by incorporating further information, such as the audio, a better classification with more classes will be possible.

In case of *action* movies, we sort them on the basis of number of shots showing fire/explosions. It is clear from

Fig.7 that the movie *The World Is Not Enough* contains more explosions/gunfire as compared to the other movies, hence not suitable for young children. Whereas, *Rush Hour* contains the least number of explosion shots.

6. Conclusions

In this paper we have proposed a method to perform a high level classification of movies into genres using the previews. In the future we plan to extend this work to analyze complete movies in order to explore the semantics from the shot level to the scene level. We also plan to utilize the grammar of movie making to present the higher level description of the entire stories.

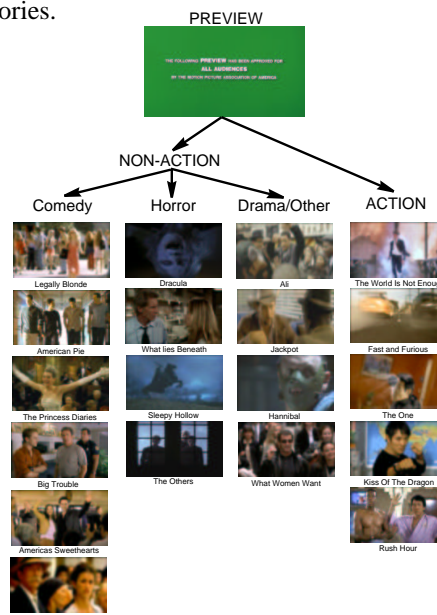


Figure 7. Final classification of Movies.

References

- [1] N. Vasconcelos, Lippman. “Towards Semantically Meaningful Feature Spaces for the Characterization of Video Content”. ICIP 1997
- [2] Nam J.,Alghoniemy M.,Tewfik A. “Audio-visual content based violent scene characterization” ICIP 1998.
- [3] Niels Hearing, “A Framework for the Design of Event Detections” Ph.D. Thesis, School of Computer Science, University of Central Florida, 1999.
- [4] B. Jahne, *Spatio-temporal Image Processing: Theory and Scientific Applications*, Springer Verlag, 1991.
- [5] Herbert Zettl, “Sight Sound Motion, Applied Media Aesthetics, Second Edition
- [6] A. F. Reynertson, “The Work of the film director”, First Edition. 1970, Hastings House.
- [7] <http://www.apple.com/trailers/>