Movie Genre Classification By Exploiting Audio-Visual Features Of Previews

Zeeshan Rasheed Mubarak Shah
Computer Vision Lab, University of Central Florida
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. We perform the initial classification into action and non-action by computing the visual disturbance feature 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. Potential applications are browsing and retrieval of videos on the Internet (video-on-demand), video libraries, and rating of the movies. This work is a step towards automatically building and updating video database, thus resulting in minimum human intervention.

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, being 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 as easy as it seems.

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. Low-level features such as color and music may be combined with the high-level domain knowledge to classify movie genre.

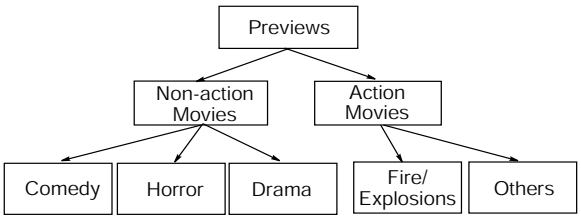


Figure 1: Flow chart showing the classes of movie genre

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 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 (Section2.1 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 Figure 1). This is done by first analyzing audio information. We find events by identifying the peaks in sound energy and test corresponding frames for the occurrence of an explosion. Sections 3 and 4 discuss the sub-classification. Section 5 presents the experimental results and finally Section 6 concludes our work.

2.1. Shot detection and Average shot length

We use 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:

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

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 $I(y,t)$, where each is defined by planes (x,t) and (y,t) for horizontal and vertical slices respectively. To find the disturbance in the scene, we evaluate the structure tensor of the slices which is expressed as:

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

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. With the help of the above equations we can solve for the angle of orientation θ 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 (for example camera translation) the gray levels of all pixels in a row change in the same direction. This results in the equal or 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.

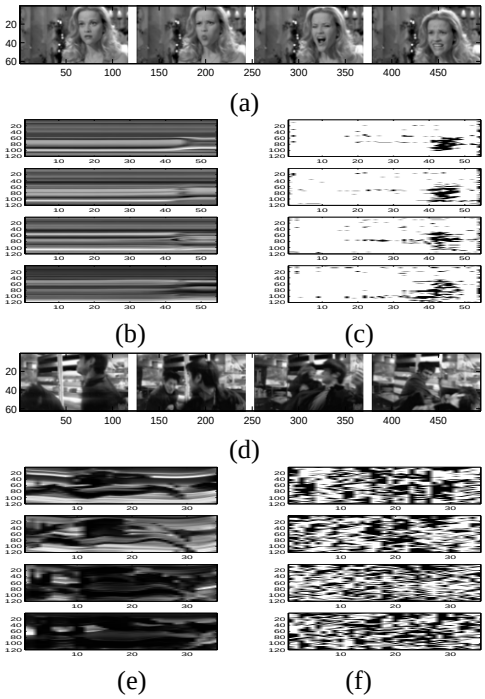


Figure 2: Plot of *Visual disturbance*, four frames of shots taken from the movie (a) *Legally Blonde* and (d) *Kiss of the Dragon*. (b) and (e) are the horizontal slices for four fixed rows of corresponding shots. (c) and (f) show active pixels (black) in corresponding slices.

We analyze the distribution of θ for each column of the horizontal slice by generating a nonlinear histogram. Based on experiments, we divide the histogram into 7 nonlinear bins with boundaries at [-90, -55, -35, -15, 15, 35, 55, 90]. The first and the last bins accumulate the higher values of θ, whereas the middle one captures the smaller values. In case of a static scene or a scene with global motion all pixels have similar value of θ and therefore they fall into one bin. On the other hand, pixels with motion other than global motion have different values of θ and fall into different bins. We locate the peak in the histogram and mark the pixels in that bin as *static*, whereas the remaining ones are marked as *moving* pixels. Next, we generate a binary mask for the whole video clip separating static from that of moving pixels. The overall *visual disturbance* is the ratio of moving pixels to the total number of the pixels in a slice.

To reduce the complexity, we use the average of disturbance of only four equally separated horizontal slices for each movie trailer as a disturbance measure. This measure is proportional to the amount of action occurring in

a shot. Figure 2 shows *visual disturbance* measure for shots of two different movies. It is clear that the density of *disturbance* is much smaller for a *non-action* shot as compared to an *action* shot. The computation of *visual disturbance* is very efficient and computationally less expensive. Our method processes only four rows per image as compared to [1] that estimates affine motion parameters for every frame and, therefore, runs very fast.

2.3. Initial classification

We have observed that *action* movies have more local motion than a *drama* or a *horror* movie which results in a larger *visual disturbance*. We have also noticed that in *action* movies shots change rapidly than in other genre like *drama* and *comedy*. We use a similar technique as suggested by [1] and plot the *visual disturbance* against average shot length. Using a linear classifier we separate *action* movies from *non-action*.

3. Sub-classification of non-action movies

The perceptual effects of color on human emotions are well known. Similarly "*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]. For example, it is less likely that a ghost would wear a rainbow color shirt, or a *comedy* movie would be shot with a dark color tone. The 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** A *high-key* lighting means that the scene has an abundance of bright light. It usually has 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** In this type the background and the part of the scene is generally predominantly dark. In *low-key* scenes, the contrast ratio is high. *Low-key* lighting being more dramatic are often used in *Film Noir* or *horror* films.

In *horror* movies shots are mostly *low-key*, especially in previews, as previews contain the most interesting scenes of the movie. 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, indicating a rich mix of colors.

(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.

Given *k* key frames, each with $m \times n$ pixels in it, we find the mean, μ, and standard deviation, σ, of gray scale distribution. For each movie, *i*, we define a quantity $\zeta_i(\mu, \sigma)$ which is the product of μ_i and σ_i, that is:

$$\zeta_i = \mu_i \cdot \sigma_i \tag{6}$$

where μ_i and σ_i are normalized. 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 therefore define two thresholds, τ_c and τ_h, and assign a category to each movie *i* based on the following criterion.

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

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

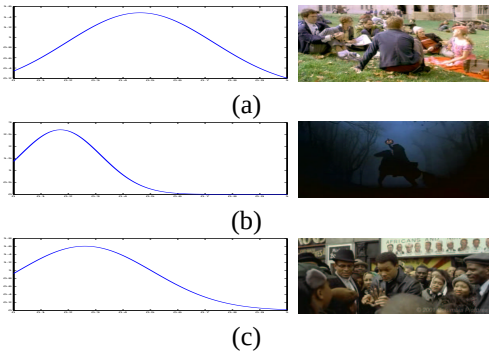


Figure 3: Average intensity histogram of key frames (a) *Legally Blonde*, a *comedy* movie (b) *Sleepy Hollow*, a *horror* movie and (c) *Ali*, an example of *drama/other*.

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 use both audio and color information to achieve this task.

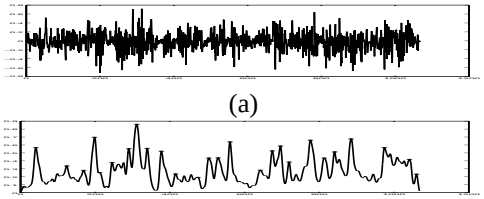
4.1. Audio analysis

Music and nonliteral sounds are often used to provide additional energy to the scene. It can quite easily describe

a condition, for example, whether a situation is stable or unstable. In case of *action* movies, the audio is always correlated with the scene. For example, fighting, explosions, etc. are mostly accompanied with a sudden change in the audio level. We, therefore, first compute the energy in the audio track using the following formula:

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

where A_i is the audio sample indexed by time i. Interval was set to 50ms. See Figure 4 for the plots of audio signal and its energy for audio track of the movie *The World Is Not Enough*. Since we are interested in the instances where the



(a)

(b)

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

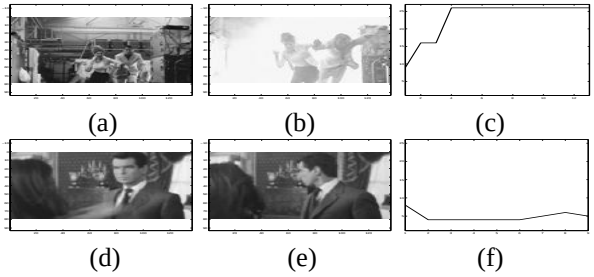
energy in audio changes abruptly, we perform a peakiness test on the energy plot. A peak is good if it is sharp and deep. That is:

$$peakiness = \left(1 - \frac{V_a + V_b}{2P}\right) \cdot \left(1 - \frac{N}{W \times P}\right) \quad (9)$$

where P is the height of the peak, V_a and V_b are the height of the valleys on either sides of the peak. W is the width of the peak and N denotes the area under the valley. Videos corresponding to the peaks above than a threshold, T_{peak}, are selected to process the corresponding video.

4.2. Fire/explosion detection

Once the occurrence of *events* in the movie are detected, we analyze the corresponding frames to detect fire and/or explosions. 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 might be considered as an explosion. Figure 5 show the detection in two candidate shots. Although several scenes had abrupt changes in audio in our experiments, our algorithm successfully differentiated between explosions and non-explosion shots.



(a) (b) (c)

(d) (e) (f)

Figure 5: Detection of *fire/explosion* in two shots. (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 successfully identified as *fire/explosion* shot whereas second shot was identified as *non-explosion* shot.

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.

Figure 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. Using the inten-

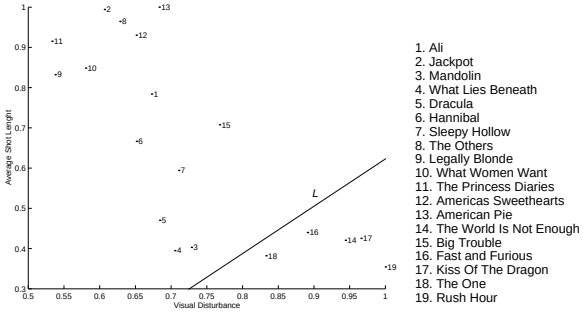


Figure 6: The distribution of Movies on the basis of *Visual Disturbance* and *Average shot length*.

sity distribution of key frames of *non-action* class, we label

Sub classification of *non-action* movies C=Comedy, H=Horror and D=Drama/Other		
1	Ali	D
2	Jackpot	D
3	Mandolin	C
4	What Lies Beneath	H
5	Dracula	H
6	Hannibal	D
7	Sleepy Hollow	H
8	The Others	H
9	Legally Blonde	C
10	What Women Want	D
11	The Princess Diaries	C
12	American Sweethearts	C
13	American Pie	C
14	Big Trouble	C

Table 1: Sub classification of *non-action* movies on the basis of key frames analysis.

Action class	
1	The World Is Not Enough
2	Fast And The Furious
3	The One
4	Kiss Of The Dragon
5	Rush Hour

Table 2: Action movies sorted according to the content of fire/explosion in their previews.

movies as *comedy, horror* and *drama/other*. Table 1 shows the sub-labels for 14 *non-action* movies. *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 the 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. Shots that were accompanied by high peaks in the audio were tested as candidates for fire/explosion movie. Table 2 shows the movies sorted based on the number of fire/explosions detected in their previews. It is clear from the above table 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. Figure 7 shows the final classification result of our experiments.

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. Our ultimate goal is to minimize the human intervention in building and updating of the video databases.

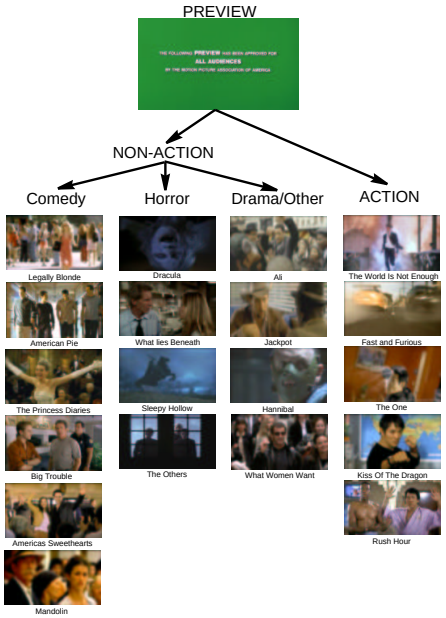


Figure 7: Classification of Movies. Note that *action* movies are sorted according to the *fire/explosion* content in the previews.

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.H.*" 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-tmporal 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/