On the Use of Computable Features for Film Classification

Zeeshan Rasheed, *Student Member, IEEE,* Yaser Sheikh, *Student Member, IEEE,* and Mubarak Shah, *Fellow, IEEE*

***Abstract*— This paper presents a framework for the classification of feature films into genres, based only on computable visual cues. We view the work as a step towards high-level semantic film interpretation, currently using low-level video features and knowledge of ubiquitous cinematic practices. Our current domain of study is the movie preview, commercial advertisements primarily created to attract audiences. A preview often emphasizes the theme of a film and hence provides suitable information for classification. In our approach, we classify movies into four broad categories: Comedies, Action, Dramas or Horror films. Inspired by cinematic principles, four computable video features (average shot length, color variance, motion content and lighting key) are combined in a framework to provide a mapping to these four high-level semantic classes. Mean shift classification is used to discover the structure between the computed features and each film genre. We have conducted extensive experiments on over a hundred film previews and notably demonstrate that low-level visual features (without the use of audio or text cues) may be utilized for movie classification. Our approach can also be broadened for many potential applications including scene understanding, the building and updating of video databases with minimal human intervention, browsing and retrieval of videos on the Internet (video-on-demand) and video libraries.**

***Index Terms*— Movie genres, previews, shot length, high-key, low-key, video-on-demand.**

I. Introduction

FILMS are a means of expression. Directors, actors, and cinematographers use this medium as a means to communicate a precisely crafted storyline. This communication operates at several levels; explicitly, with the delivery of lines by the actors, and implicitly, with the background music, lighting, camera movements and so on. Directors often follow well-established rules, commonly called 'film grammar' in literature, to communicate these concepts. Like any natural language, this grammar has several dialects, but is more or less universal. This fact in film-making (as compared to arbitrary video data) suggests that knowledge of cinematic principles can be exploited effectively for the understanding of films. To interpret an idea using the grammar, we need to first understand the symbols, as in natural languages, and second, understand the rules of combination of these symbols to represent concepts. Daniel Arijon, a famous name in film literature, writes, "All the rules of film grammar have been on the screen for a long time. They are used by filmmakers as far apart geographically and in style as Kurosawa in Japan, Bergman in Sweden, Fellini in Italy and Ray in India. For them, and countless others this common set of rules is used to solve specific problems presented by the visual narration of a story", [3], page 4.

In order to exploit knowledge of these ubiquitous techniques, it is necessary to be able to relate the symbols of film grammar to *computable video features*. Computable video features, as the name suggests, are defined as any statistic of the available video data. Since the relationship between film grammar symbols and high-level film semantics is known, if we are able to find computable representations of these symbols, the problem of film classification can be favorably posed. Unfortunately, not all the symbols of film grammar can be well-represented in terms of a statistic. For instance, how does one compute the irony in a scene? It is immediately evident that *high-level* symbols like emotion, irony, or gestures are difficult to represent as statistics. On the other hand, *low-level* symbols like lighting, shot length and background music are far easier to represent. It should also be noted that low-level symbols correspond to the implicit communication that the director uses, and incidently are also the type of symbols that have the most established techniques. Audiences too become 'trained' to interpret low-level symbols in a certain way, as is evidenced by feelings of expectation associated with silence, or feelings of fear associated with dim-lighting. These ideas are investigated in depth in [24], [3].

Films constitute a large portion of the entertainment industry. Every year about 4,500 films are released around the world, which correspond to approximately 9,000 hours of video, [28]. While it is feasible to classify films at the time of production, classification at finer levels, for instance classification of individual scenes, would be a tedious and substantial task. Currently, there is a need for systems to extract the 'genre' of scenes in films. Application of such scene-level classification would allow departure from the prevalent system of *movie* ratings to a more flexible system of *scene* ratings. For instance, a child would be able to watch movies containing a few scenes with excessive violence, if a pre-filtering system can prune out scenes that have been rated as violent. Such semantic labelling of scenes would also allow far more flexibility while searching movie databases. For example, automatic recommendation of movies based on personal preferences could help a person choose a movie, by executing a scene level analysis of previously viewed movies. While the proposed method does not actually achieve scene classification, it provides a suitable framework for such work.

Some justification must be given for the use of previews for the classification of movies. Since movie previews are primarily commercial advertisements, they tend to emphasize the theme of the movie, making them particularly suited for the task of genre classification. For example, previews of

Manuscript received January 20, 2002; revised November 18, 2002.
The authors are with the Department of Computer Science, University of Central Florida, Orlando, Florida 32816-2362

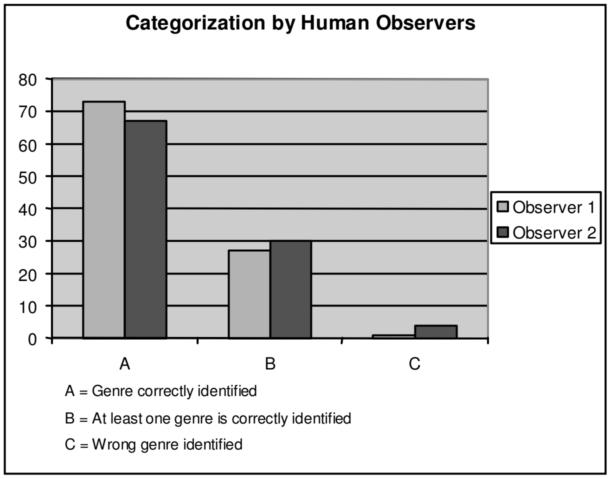


Fig. 1. Genre classification by Human Observers with respect to the ground truth obtained from the IMDB and Apple web-site. Observers were asked to categorize films into four genres based on their previews

action movies inevitably contain shots of fights, chases and sometimes crashes, explosions and gunfire. Exceptions exist, of course, and in order to strengthen the claim that high-level semantic classification based on previews is *possible*, we conducted human evaluations of our data set consisting of over a hundred film previews and compared the results with the ground truth obtained from IMDB (Internet Movie Database, [4] and Apple web-site, [2]). Two observers were asked to watch the previews and classify each movie into the four genres. Both observers managed to identify at least one of the genres in the ground truth, for practically all the movies. Ignoring bias of prior knowledge, what the experiment suggested was that classification based on movie previews is, at the very least, *possible*. The results of the evaluation are displayed in Figure 1. In conclusion, we present a framework for genre classification based on four computed features, average shot length, color variance, motion content and lighting key, from film previews. It is noteworthy that in this work inferences are made using visual features *only*, and no audio or textual information is used. Furthermore, since both previews and scenes are composed of several shots, this framework can be suitably extended for applications of scene classification.

The rest of the paper is organized as follows. Related work is discussed in Section II. In Section III, we present the computable video features that are used for classification in our work. Section IV details the use of mean shift classification as a clustering approach in our application. A discussion of the results is presented in Section V, followed by conclusions in Section VI.

II. Related Work

One of the earliest research efforts in the area of video categorization and indexing was the Informedia Project [13] at Carnegie Mellon University. It spearheaded the effort to segment and automatically generate a database of news broadcasts every night. The overall system relied on multiple low-level cues, like video, speech, close-captioned text and other cues. However, there are a few approaches which deal with higher-level semantics, instead of using low-level feature matching as the primary indexing criteria. Work by Fischer *et al* in [11] and a similar approach by Truong *et al* in [26], distinguished between newscasts, commercials, sports, music videos and cartoons. The feature set consisted of scene length, camera motion, object motion and illumination. These approaches were based on training the system using examples and then employing a decision tree to identify the genre of the video.

Content based video indexing also constitutes a significant portion of the work in this area. Chang *et al* [7] developed an interactive system for video retrieval. Several attributes of video such as color, texture, shape and motion were computed for each video in the database. The user was required to provide a set of parameters for attributes of the video that was being searched for. These parameters were compared with those in the database using a weighted distance formula for the retrieval. A similar approach has also been reported by Deng *et al* [9].

The use of Hidden Markov Models has been very popular in the research community for video categorization and retrieval. Naphade *et al* [19] proposed a probabilistic framework for video indexing and retrieval. Low-level features were mapped to high-level semantics as probabilistic multimedia objects called *multijects*. A Bayesian belief network, called a *multinet*, was developed to perform semantic indexing using Hidden Markov Models. Some other examples that make use of probabilistic approaches are [29], [10], [6]. Qian *et al* also suggested a semantic framework for video indexing and detection of events. They presented an example of hunt detection in videos, [23].

A large amount of research work on video categorization has also been done in the compressed-domain using MPEG-1 and MPEG-2. The work in this area utilizes the extractable features from compressed video and audio. Although the compressed information may not be very precise, it avoids the overhead of computing features in the pixel domain. Kobla *et al* [16] used DCT coefficients, macroblock and motion vector information of MPEG videos for indexing and retrieval. Their proposed method was based on *Query-by-Example* and found the spatial (DCT coefficients) and temporal (motion) similarities among the videos using *FastMap*. The methods proposed in [30], [20] are a few more examples which also work on compressed video data. Lu *et al* [17] applied an HMM based approach in the compressed domain and promising results were presented. Recently, the MPEG-7 community has focused on video indexing by using embedded semantic descriptors, [5]. However, the standardization of MPEG-7 is currently under development and the content-to-semantic interpretation for retrieval of videos is still an open question for the research community.

Specific to film classification, Vasconcelos *et al* proposed a feature-space based approach in [27]. In this work, two features of the previews, average shot length and shot activity, were used. In order to categorize movies they used a linear classifier in the two-dimensional feature space. An extension of their approach was presented in Nam *et al*, [18], which identified violence in previews. They attempted to detect violence using audio and color matching criteria. One problem with these existing approaches in film classification is the crude structure that is imposed while classifying data (in the form of the linear classifier). In our work, we adopt a non-parametric approach, using mean shift clustering. Mean shift clustering has been shown to have excellent properties for clustering real data. Furthermore, we exploit knowledge of cinematic principles, presenting four computable features for the purposes of classification. We believe that the *extendibility* of the proposed framework to include new, possibly higher-level features is an important aspect of the work. Since the approach discovers the structure of the mapping between features and classes autonomously, the need to handcraft rules of classification is no longer required.

III. Computable Video Features

In this paper, we present the problem of semantic classification of films within the feature-space paradigm. In this paradigm, the input is described through a set of features that are likely to *minimize* variance of points within a class and *maximize* variance of points across different classes. A parametric representation of each feature is computed and is mapped to a point in the multidimensional space of the features. Of course, the performance depends heavily on the selection of appropriate features. In this section, we present four computable features that provide good discrimination between genres. An (arguably) comprehensive list of genres can be found at `http://us.imdb.com/Sections/Genres/`, which enumerates them as Action, Adventure, Animation, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film Noir, Horror, Musical, Mystery, Romance, Science Fiction, Short, Thriller, War and Western. From this list we identified four *major* genres, namely Action, Comedy, Horror and Drama. There are two reasons for this choice. Firstly, these genres represent the majority of movies currently produced, and most movies can be classified, albeit loosely, into at least one of these major genres. Secondly, we have selected these four genres since it is between these genres that low-level discriminant analysis is most likely to succeed. In other words, it is in these genres that we propose correlation exists between computable video features and their respective genres. However, the data set itself was not pre-screened to fit specifically into one of these genres, as the subsequent results will show, many movies fit more than one of the categories. Rather than espouse individual genre classification, we acknowledge the fact that a film may correctly be classified into *multiple* genres. For instance, many Hollywood action films produced these days have a strong element of comedy as well. In the remainder of this section, we discuss the four features that are employed for classification, namely average shot length, shot motion content, lighting key and color variance.

A. Shot Detection and Average Shot Length

The first feature we employ is the average shot length. This feature was first proposed by Vasconcelos in [27]. The average shot length as a feature represents the tempo of a scene. The director can control the speed at which the audience's attention is directed by varying the tempo of the scene, [1]. The average shot length provides an effective measure of the tempo of a scene, and the first step in its computation is the detection of shot boundaries. A shot is defined as a sequence of frames taken by a single camera without any major change in the color content of consecutive images. Techniques based on color histogram comparison have been found to be robust and are used by several researchers for this purpose. In our approach, we extend the algorithm reported in [12] for the detection of shot boundaries using HSV color histogram intersection.

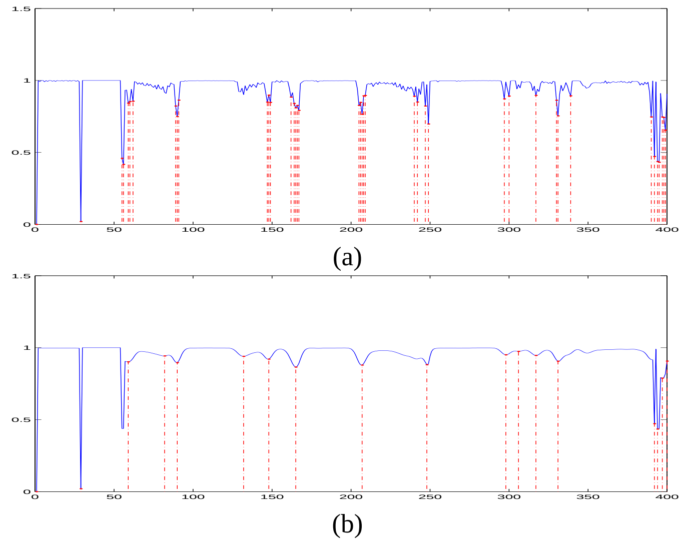


Fig. 2. (a) Shot detection using fixed threshold method for a segment of the movie trailer of *Red Dragon*. Only first 400 frames are shown for convenient visualization. There are 17 shots identified by a human observer. (a) Fixed threshold method. Vertical lines indicate the detection of shots. Number of shots detected: 40, Correct: 15, False positive: 25, False negative: 2 (b) Shots detected by proposed method. Number of shots detected: 18, Correct: 16, False positive: 2, False negative: 1

Each histogram consists of 16 bins; 8 for the hue, 4 for the saturation and 4 for the value components of the HSV color space. Let $S(i)$ represent the intersection of histograms H_i and H_{i-1} of frames i and $i-1$ respectively. That is:

$$S(i) = \sum_{j \in allbins} min(H_i(j), H_{i-1}(j)). \tag{1}$$

The magnitude $S(i)$ is often used as a measure of shot boundary in related works. The values of i where $S(i)$ is less than a fixed threshold are assumed to be the shot boundaries. This approach works quite well (see [12]) if the shot change is abrupt and there are no shot transition effects (wipes, dissolves etc.) Previews are generally made with a variety of shot transition effects. We have observed that the most commonly used transition effect in previews is a *dissolve* in which several frames of consecutive shots overlap. Applying a fixed threshold to $S(i)$ when the shot transition occurs with a *dissolve* generates several outliers because consecutive frames differ from each other until the shot transition is completed.

To improve the accuracy, an iterative smoothing of the one dimensional function S is performed first. We have adapted the algorithm proposed by Perona *et al* [22] based on anisotropic diffusion. This is done in the context of scale-space. S is

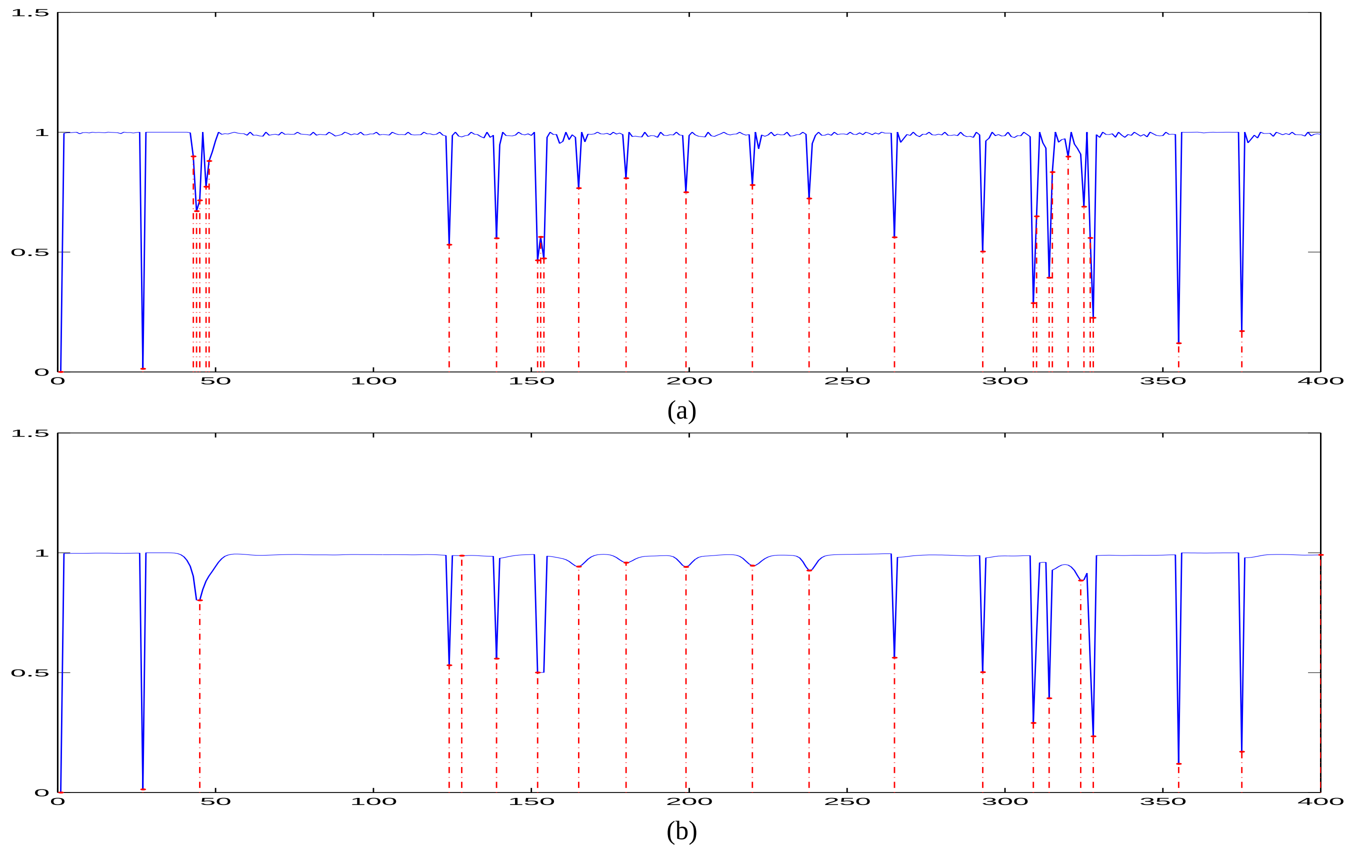


Fig. 3. (a) Shot detection using the fixed threshold method for a segment of the preview of *Road Trip*. Only first 400 frames are shown for convenient visualization. There are 19 shots identified by a human observer. (a) Fixed threshold method. Vertical lines indicate the detection of shots. Number of shots detected: 28, Correct: 19, False positive: 9, False negative: 0. (b) Shots detected by proposed method. Number of shots detected: 19, Correct: 19, False positive: 0, False negative: 0.

smoothed iteratively using a Gaussian kernel such that the variance of the Gaussian function varies with the signal gradient. Formally,

$$S^{t+1}(i) = S^t(i) + \lambda \left[c_E \cdot \nabla_E S^t(i) + c_W \cdot \nabla_W S^t(i) \right], \quad (2)$$

where t is the iteration number and $0 < \lambda < 1/4$ with:

$$\begin{aligned} \nabla_E S(i) &\equiv S(i+1) - S(i), \\ \nabla_W S(i) &\equiv S(i-1) - S(i). \end{aligned} \quad (3)$$

The condition coefficients are a function of the gradients and are updated for every iteration,

$$\begin{aligned} c_E^t &= g\left(\mid \nabla_E S^t(i) \mid \right), \\ c_W^t &= g\left(\mid \nabla_W S^t(i) \mid \right), \end{aligned} \quad (4)$$

where $g(\nabla_E S) = e^{-(\frac{|\nabla_E|}{k})^2}$ and $g(\nabla_W S) = e^{-(\frac{|\nabla_W|}{k})^2}$. In our experiments, the constants were set to $\lambda = 0.1$ and $k = 0.1$. Finally, the shot boundaries are detected by finding the local minima in the smoothed similarity function S. Thus, a shot boundary will be detected where two consecutive frames will have minimum color similarity. This approach reduces the false alarms produced by the fixed threshold method.

Figure 2 presents a comparison between the two methods, (a) using a fixed threshold method and (b) using the proposed method. The similarity function S is plotted against the frame numbers. Only the first 400 frames are shown for convenient visualization. There are several outliers in (a) because gradually changing visual contents from frame to frame (the dissolve effect) are detected as a shot change. For instance, there are multiple shots detected around frame numbers 50, 150 and 200. However, in (b), a shot is detected when the similarity between consecutive frames is minimum. Compare the detection of shots with (a). Figure 3 shows improved shot detection for the preview of *Road Trip*. See Table I that lists precision and recall of shot detection for some of the trailers in the data set.

The average shot length is then computed for each preview. This feature is directly computed by dividing the total number of frames by the total number of shots in the preview (the statistical mean). Our experiments show that slower paced films such as dramas have larger average length as they have many dialogue shots, whereas action movies appear to have shorter shot lengths because of rapidly changing shots. Each detected shot is represented by a key-frame to analyze shot's color attributes. We use the middle frame of each shot as the key-frame.

B. Color Variance

Zettl observes in, [31], " The expressive quality of color is, like music, an excellent vehicle for establishing or intensifying the mood of an event." In this work, we are interested in

TABLE I
EXAMPLES OF SHOT DETECTION RESULTS IN SOME PREVIEWS IN THE DATA SET.

Shot Detection Results		
Movie	**Recall**	**Precision**
24 Hours Party People	0.96	0.84
Ali	0.85	0.91
American Pie	0.99	0.98
Americas Sweethearts	0.95	0.92
Big Trouble	0.89	0.92
Dracula 2000	0.95	0.96
The Fast and the Furious	0.93	0.87
Hannibal	0.94	0.86
The Hours	0.88	0.99
Jackpot	0.96	0.93
Kiss Of The Dragon	0.97	0.90
Legally Blonde	0.98	0.96
Mandolin	0.91	0.95
Red Dragon	0.96	0.91
Road Trip	1.00	0.99
Rush Hour	0.94	0.91
Sleepy Hollow	0.96	0.89
Stealing Harvard	0.98	0.95
The One	0.95	0.86
The Others	0.91	0.95
The Princess Diaries	0.90	0.88
The World Is Not Enough	0.96	0.83
The Tuxedo	0.98	0.91
What Lies Beneath	0.97	0.97
What Women Want	0.96	0.94

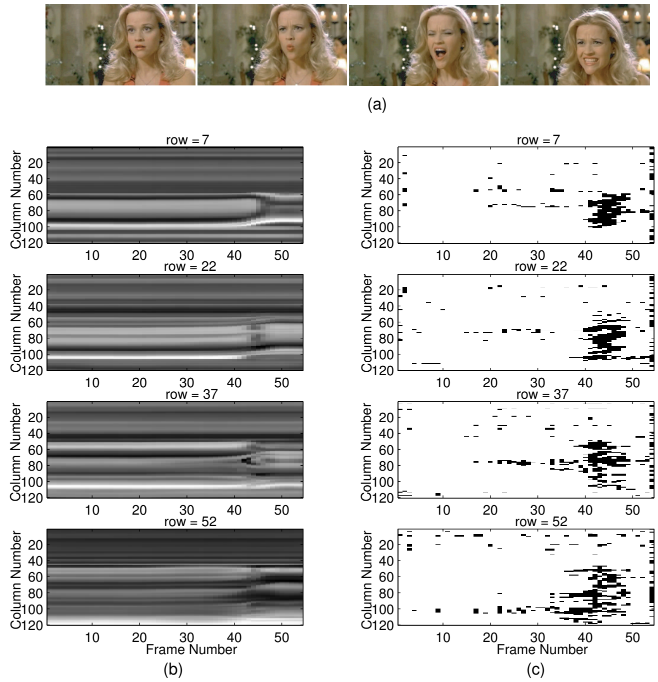


Fig. 4. Plot of *Visual disturbance*. (a) Four frames of shots taken from the preview of *Legally Blonde*. (b) Horizontal slices for four fixed rows of a shot from the preview. Each column in the horizontal slice is a row of image. (c) Active pixels (black) in corresponding slices.

exploiting the variance of color in a clip *as a whole* to discriminate between genres of a film. Intuitively, the variance of color has a strong correlational structure with respect to genres, as it can be seen, for instance, that comedies tend to have a large variety of bright colors, whereas horror films often adopt only darker hues. Thus, in order to define a computable feature two requirements have to be met. First, a feature has to be defined that is *global* in nature, and second, distances in the color space employed should be perceptually uniform. We employ the CIE Luv space, which was designed to approach a perceptually uniform color space. To represent the variety of color used in the video we employ the generalized variance of the Luv color space of each preview as a whole. The covariance matrix of the multi-variate vector (three dimensional in our case) is defined as,

$$\rho = \begin{bmatrix} \sigma_L^2 & \sigma_{Lu}^2 & \sigma_{Lv}^2 \\ \sigma_{Lu}^2 & \sigma_u^2 & \sigma_{uv}^2 \\ \sigma_{Lv}^2 & \sigma_{uv}^2 & \sigma_v^2 \end{bmatrix}. \quad (5)$$

The *generalized variance* is obtained by finding the determinant of Equation 5,

$$\sum = \det(\rho) \quad (6)$$

This feature is used as a representation of the color variance. All key-frames present in a preview are used to find this feature.

C. Motion Content

The *visual disturbance* of a scene can be represented as the motion content present in it. The motion content represents the amount of activity in a film. Obviously, action films would have higher values for such a measure, and less visual disturbance would be expected for dramatic or romantic movies. To find *visual disturbance*, an approach based on the structural tensor computation is used which was introduced in [14]. The frames contained in a video clip can be thought of as a volume obtained by considering 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, the structure tensor of the slices is evaluated, which is expressed as,

$$\mathbf{\Gamma} = \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} = \begin{bmatrix} \sum_w H_x^2 & \sum_w H_x H_t \\ \sum_w H_x H_t & \sum_w H_t^2 \end{bmatrix}, \quad (7)$$

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:

$$R \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} R^T = \begin{bmatrix} \lambda_x & 0 \\ 0 & \lambda_t \end{bmatrix}, \quad (8)$$

where λ_x and λ_y are the eigenvalues and R is the rotation matrix. With the help of the above equations we can solve for the orientation angle θ as

$$\theta = \frac{1}{2} tan^{-1} \frac{2J_{xt}}{J_{xx} - J_{tt}}. \quad (9)$$

When there is no motion in a shot, θ is constant for all pixels. With global motion (e.g. camera translation) the gray levels of all pixels in a row change in the same direction.

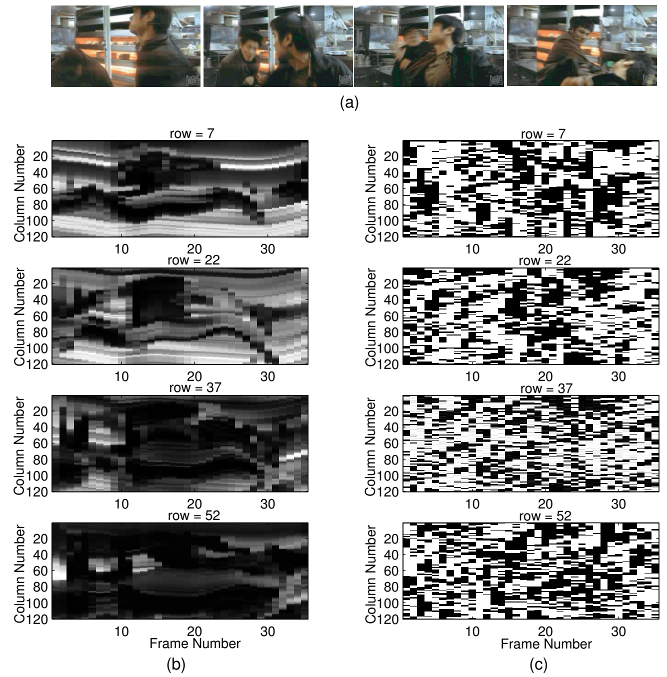


Fig. 5. Plot of *Visual disturbance*. (a) Four frames of shots taken from the preview of *Kiss of the Dragon*. (b) Horizontal slices for four fixed rows of a shots from the preview. Each column in the horizontal slice is a row of image. (c) Active pixels (black) in corresponding slices.

This results in equal or similar values of θ. However, in the case of local motion, pixels that move independently will have different orientations. This can be used to label each pixel in a column of a slice as a moving or a non-moving pixel.

The distribution of θ for each column of the horizontal slice is analyzed by generating a nonlinear histogram. Based on experiments, the histogram is divided into 7 nonlinear bins with boundaries at [-90, -55, -35, -15, 15, 35, 55, 90] degrees. The first and the last bins accumulate the higher values of θ, whereas the middle one captures the smaller values. In 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. The peak in the histogram is located and the pixels in the corresponding bin are marked as *static*, whereas the remaining ones are marked as *active* pixels. Next, a binary mask for the whole video clip is generated separating static pixels from active ones. The overall motion content is the ratio of moving pixels to the total number of pixels in a slice. Figures 4 and 5 show motion content measure for two shots. Figure 4 is a dialogue shot taken from the movie *Legally Blonde*. On the other hand, Figure 5 is a shot taken from a fight scene with high activity. Compare the density of moving pixels (black pixels in (c)) of both figures. It should be noted that the density of motion is much smaller for a non-action shot as compared to an action shot.

D. Lighting Key

In the hands of an able director, lighting is an important dramatic agent. Generations of film-makers have exploited luminance to evoke emotions, using techniques that are well studied and documented in cinematography circles, [31]. A deliberate relationship exists, therefore, between the lighting and the genre of a film.

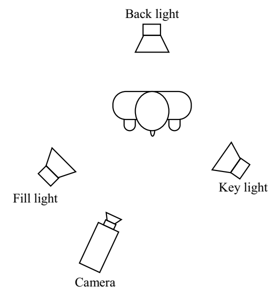


Fig. 6. Positioning of lights in a three-point lighting setup.

In practice, movie directors use multiple light sources to balance the amount and direction of light while shooting a scene. The purpose of using several light sources is to enable a specific portrayal of a scene. For example, how and where shadows appear on the screen is influenced by maintaining a suitable proportion of intensity and direction of light sources. Lighting can also be used to direct the attention of the viewer to certain area of importance in the scene. It can also affect viewer's feeling directly regardless of the actual content of the scene. Reynertson comments on this issue, "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." [24], page 107. In other words, lighting is an issue not only of enough light in the scene to provide good exposure, but of light and shade to create a dramatic effect, consistent with the scene. In a similar vein, Wolf Rilla says "All lighting, to be effective, must match both mood and purpose. Clearly, heavy contrasts, powerful light and shade, are inappropriate to a light-hearted scene, and conversely a flat, front-lit subject lacks the mystery which back-lighting can give it." [25] page 96.

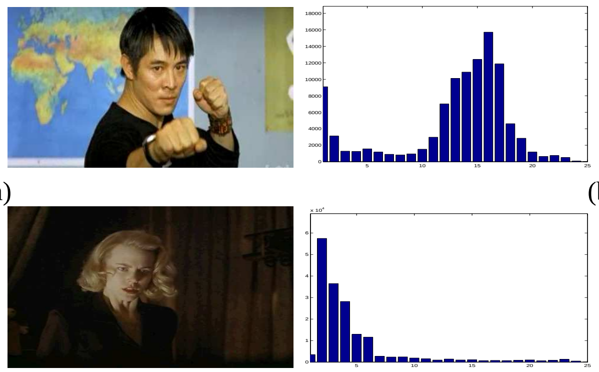
There are numerous ways to illuminate a scene. One of the commonly used methods in the film industry is called *Three Point Lighting*. As the name implies, this style uses three main light sources:

Key-light: This is the main source of light on the subject. It is the source of greatest illumination.

Back-light: This source of light helps emphasize the contour of the object. It also separates it from a dark background.

Fill-light: This is a secondary illumination source which helps to soften some of the shadows thrown by the key-light and back-light.

Figure 6 shows how the light sources are placed with respect to the camera and the subject. With different proportions of intensity of each source, movie directors *paint* the scene with light and typify the situation of the scene. Thus, within the design phase of a scene there is deliberate correlation between scene context and the lighting of the scene. In film literature, two major lighting methods are used to establish such a relation between the context and the mood of the viewer, called *low-key lighting* and *high-key lighting*.



(a) (b) (c) (d)

Fig. 7. Distribution of gray scale pixel values in (a) *high-key* shot (b) histogram and (c) *low-key* shot (d) histogram.

High-key lighting: *High-key* lighting means that the scene has an abundance of bright light. It usually has less contrast and the difference between the brightest light and the dimmest light is small. Practically, this configuration is achieved by maintaining a low *key-to-fill* ratio i.e. a low contrast between dark and light. *High-key* scenes are usually contain action or are less dramatic. As a result, comedies and action films typically have high-key lighting[31], page 32.

Low-key lighting: In this type, the background of the scene is generally predominantly dark. In *low-key* scenes, the contrast ratio is high. *Low-key* lighting is more dramatic and is often used in *film noir* or *horror* films.

Many algorithms exist that compute the position of a light source in a given image [21]. If the direction and intensity of the light sources are known, the *key* of the image can be easily deduced, and some higher-level interpretation of the situation can be elicited. Unfortunately, for general scenes, of the nature usually encountered in films, assumptions typically made in existing algorithms are violated, for example, single light source or uniform Lambertian surface. However, it is still possible to compute the *key* of lighting using simple computation. The brightness value of pixels in an image vary proportionally with the scene illumination and the surface properties of the observed object. Hence a *high-key* shot, which is more illuminated than a *low-key* shot, contains a higher proportion of bright pixels. On the other hand, a *low-key* frame contains more pixels of lower brightness. This simple property has been exploited here to distinguish between these two categories. Figure 7 shows the distribution of brightness values of *high* and *low* key shots. It can be roughly observed from the figure that for low-key frames, both the mean and the variance are low, whereas for high key frames the mean and variance are both higher. Thus, for a given key frame, i, with $m \times n$ pixels in it, we find the mean, μ, and standard deviation, σ, of the value component of the HSV space. The value component is known to correspond to brightness. A scene lighting quantity $\zeta_i(\mu, \sigma)$ is then defined as a measure of the lighting key of a frame,

$$\zeta_i = \mu_i \cdot \sigma_i. \tag{10}$$

In *high-key* frames, the light is well distributed which results in higher values of standard deviation and the mean. Whereas, in *low-key* shots, both μ and σ are small. This enables us to formally interpret a higher-level concept from low-level information, namely the key of the frame. In general, previews contain many important scenes from the movie. Directors pick the shots that emphasize the theme and put them together to make an interesting preview. For example, in *horror* movies, the shots are mostly *low-key* to induce fear or suspense. On the other hand, *comedy* movies tend to have a greater number of *high-key* shots, since they are less dramatic in nature. 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 due to wider distribution of gray levels.

IV. MEAN SHIFT CLASSIFICATION

Thus far, we have discussed the relevance of various low-level features of video data, implying a formulation based on feature-space analysis. The analysis of the feature-space itself is a critical step that determines both the effectiveness and the practicality of the method. Even with a highly discriminating feature space, if the analysis is rule-based or imposes an unwarranted structure on the data (e.g. linear classifiers, elliptical shape etc.) the possibility of extending or deploying the work becomes suspect. Extendibility, in particular, is a central aspect of this work, as we ultimately envision an inter-dependent, low-to-high level analysis towards semantic understanding. Although a multitude of techniques exist for the analysis of feature spaces, (see [15] for a recent survey), most are unsuited for the analysis of real data. In contrast, the mean shift procedure has been shown to have excellent properties for clustering and mode-detection with real data. An in-depth treatment of the mean shift procedure can be found in [8]. Two salient aspects of mean shift based clustering that make it suited to this application is its ability to automatically detect the number of clusters, and the fact that it is non-parametric in nature (and as a result does not impose regular structure during estimation). Since the four-dimensional feature space is composed of the Lighting-key, Average Shot Length, Motion Content and Color Variance we employ a joint domain representation. To allow separate bandwidth parameters for each domain, the product of four univariate kernels define the multivariate kernel, that is

$$K(\mathbf{x}) = \frac{C}{h_1 h_2 h_3 h_4} \prod_{i=1}^{4} k\left(\frac{x_i^2}{h_i}\right), \tag{11}$$

where x_i, $i = 1$ to 4, corresponds to the average shot length, color variance, motion content and lighting key respectively, h_i are their corresponding bandwidth parameters, and C is a normalization constant. A normal kernel is used, giving a mean shift vector of

$$\mathbf{m}_{n,N}(\mathbf{y}_j) = \mathbf{y}_{j+1} - \mathbf{y}_j = \frac{\sum_{i=1}^{4} \mathbf{x}_i \exp\left(\| \frac{x - x_i}{h_i} \|^2\right)}{\sum_{i=1}^{4} \exp\left(\| \frac{x - x_i}{h_i} \|^2\right)} - \mathbf{y}_j. \tag{12}$$

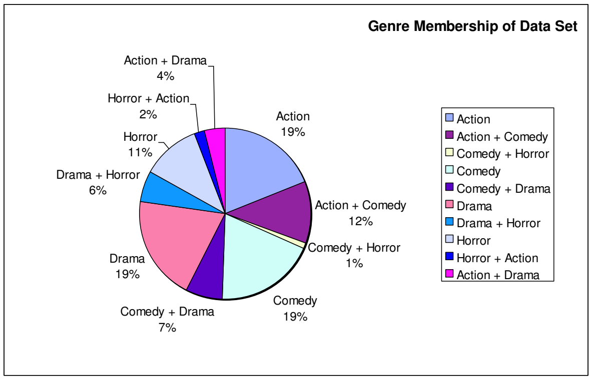


Fig. 8. Genre Membership of Data Set. It should be noted that some films have a second genre, as well.

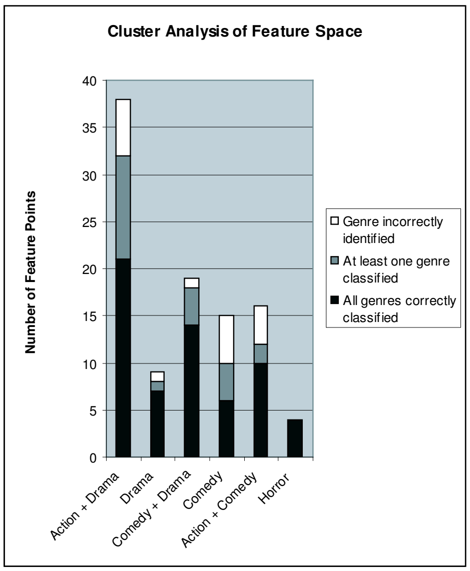


Fig. 9. Cluster Analysis of Feature Space. Six clusters are observed in the data and each cluster is classified by its dominating genre.

Mean shift clustering provides a means to analyze the feature space without making arbitrary assumptions, and lets the data define the probabilities of membership, so to speak. This formulation enables us to examine how well the computable features discriminate between the high-level labels known *a priori*. As a result, an exemplar based labelling system is also facilitated, since if there are consistent clusters and the label of one within the cluster is known, labels can be assigned to the rest.

V. Results and Discussion

We have conducted extensive experiments on just over a hundred film previews. These previews were obtained from the Apple web-site, [2]. For each preview, video tracks were analyzed at a frame rate of 12fps. The results of our experiments indicate interesting structure within the feature space, implying that a mapping does indeed exist between high-level classification and low-level computable features. We identified four major genres, namely Action, Comedy, Horror and Drama. We will first present our data set and the associated ground truth, followed by experimental results and discussion.

To investigate the structure of our proposed low-level feature space we collected a data set of 101 film previews, the ground truth of which is graphically displayed in Figure 8. As mentioned earlier, classifying movies into binary genres (as opposed to mixed genres) is unintuitive, since modern cinema often produces films with more than one theme (presumably for both aesthetic and commercial reasons). Thus, we study multiple memberships both within the ground truth and the output of the proposed method. We performed mean shift classification over all the data points in the feature space, and studied the statistics of each cluster that formed. In the following discussion, we refer to the ground truth genres as *labels* and the cluster genres as *classes*.

The data formed 6 clusters in the four-dimensional feature space, the analysis of which are displayed in Figure 9. Each cluster was assigned the label of the 'dominating genres' in the cluster. We analyzed each cluster formed, counting number of films (1) with all genres correctly identified (2) at least one genre correctly identified and (3) no genre correctly identified. The first (and the largest) cluster that was identified was the Action-Drama cluster, with 38 members. Although, only five movies were labelled Action-Dramas in the ground truth, *all* five appeared within this cluster. Moreover, the remaining points within this cluster were composed of ten films labelled as action films, and six films labelled as dramas. Eleven films with at least one genre labelled as drama or action were also observed in this cluster. The majority of the outliers (five out of six) came from the Horror genre.

The dominating genre in the second cluster was drama, with nine members. Nineteen films were labelled dramas in the ground truth, and eight of them were classified in this cluster. Only one outlier was observed within this cluster, Darkness Falls, which was labelled Horror in the ground truth. The third cluster was classified as Comedy and Drama, with nineteen members. Seven films were initially labelled as comedic dramas in the ground truth, and four of these seven were classified in this cluster. The cluster contained eight films labelled as comedies and two films labelled as dramas. The only outlier was the horror film, Session 9. The fourth cluster, classified as comedy, contained the highest percentage of outliers. Of the fifteen films in the cluster, six were labelled comedies, four had at least one genre labelled as comedy, and five were incorrectly identified. The fifth cluster was classified as Action and Comedy and had a population of sixteen. Four films in this cluster were labelled as Action Comedies, five were action movies, and one was a comedy. In the last cluster, classified as Horror, we had four horror films grouped together. This small cluster can be seen as the only successful cluster of horror films, showing that while our features are not sufficiently discriminating for *all* horror films, it captures *some* of the structure that exists. Since our feature space is four dimensional, we cannot visually display the clustering. In order to give the reader some feeling of the results, Figure 10 displays the 'profile' of each feature. The films are indexed according to their association with each cluster. See Figure 11 for thumbnails of some the film previews in the data set associated with their classes.

The total number of outliers in the final classification was

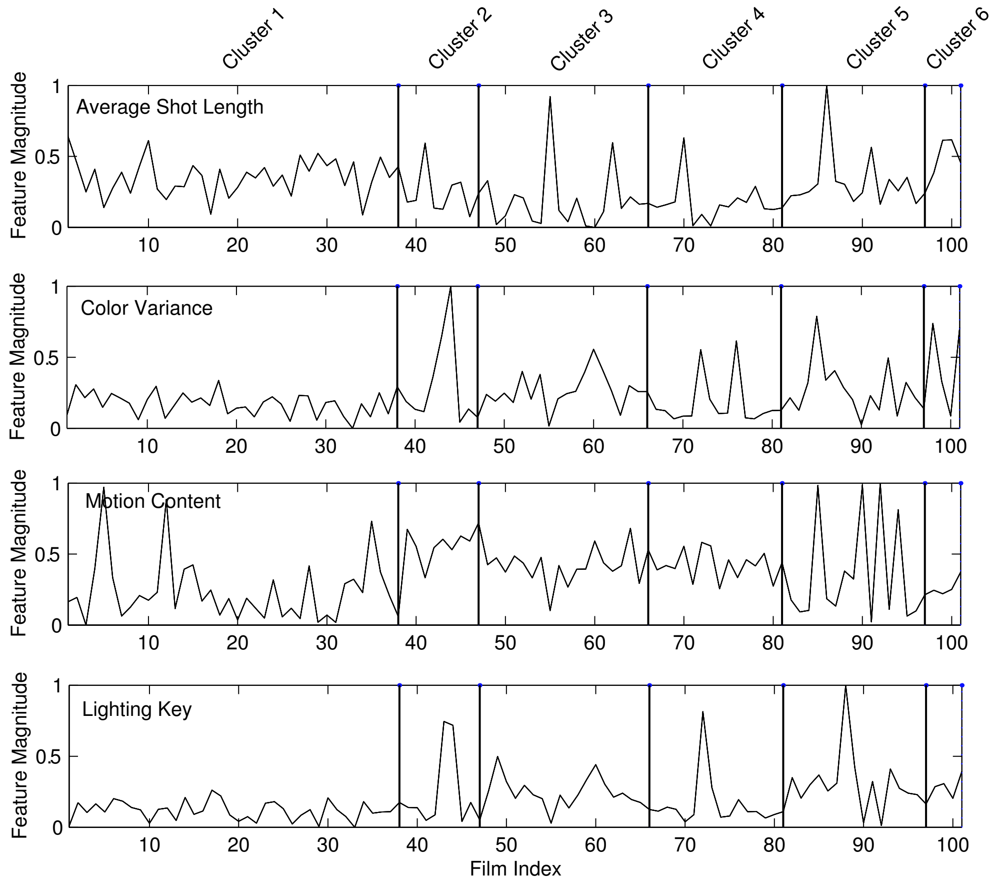


Fig. 10. Profiles of each Feature. The films are indexed according to their association with each cluster.

17 (out of 101). While this number cannot be interpreted as an 83% genre classification accuracy, it strongly supports the claim that a mapping exists between low-level video features and high-level film classes, as predicted by film literature. Thus, this domain provides a rich area of study, from the extension and application of this framework to scene classification, to the exploration of higher-level features. And as the entertainment industry continues to burgeon, the need for efficient classification techniques is likely to become more pending, making automated film understanding a necessity of the future.

VI. Conclusion

In this paper, we have proposed a method to perform high-level classification of previews into genres using low-level computable features. We have demonstrated that combining visual cues with cinematic principles can provide powerful tools for genre categorization. Classification is performed using mean shift clustering in the four dimensional feature space of Average Shot Length, Color Variance, Motion Content and the Lighting Key. We discussed the clustering thus obtained and its implications in the Results section. We plan to extend this work to analyze complete movies and to explore the semantics from the shot level to the scene level. We also plan to utilize the grammar of movie making to discover the higher level description of the entire stories. Furthermore, we are interesting in developing computable features for mid-level and high-level information, as an inter-dependent multi-level analysis is envisaged. The ultimate goal is to construct an autonomous system capable of understanding the semantics and structure of films, paving the way for many ‘intelligent’ indexing and post-processing applications.

Acknowledgment

We would like to thank Lisa Spencer for her useful comments and corrections.

References

[1] B. Adams, C. Dorai, and S. Venkatesh. Towards automatic extraction of expressive elements from motion pictures: tempo. In *IEEE International Conference on Multimedia and Expo*, pages 641–644, 2000.

[2] Apple. http:// www.apple.com/ trailers/.

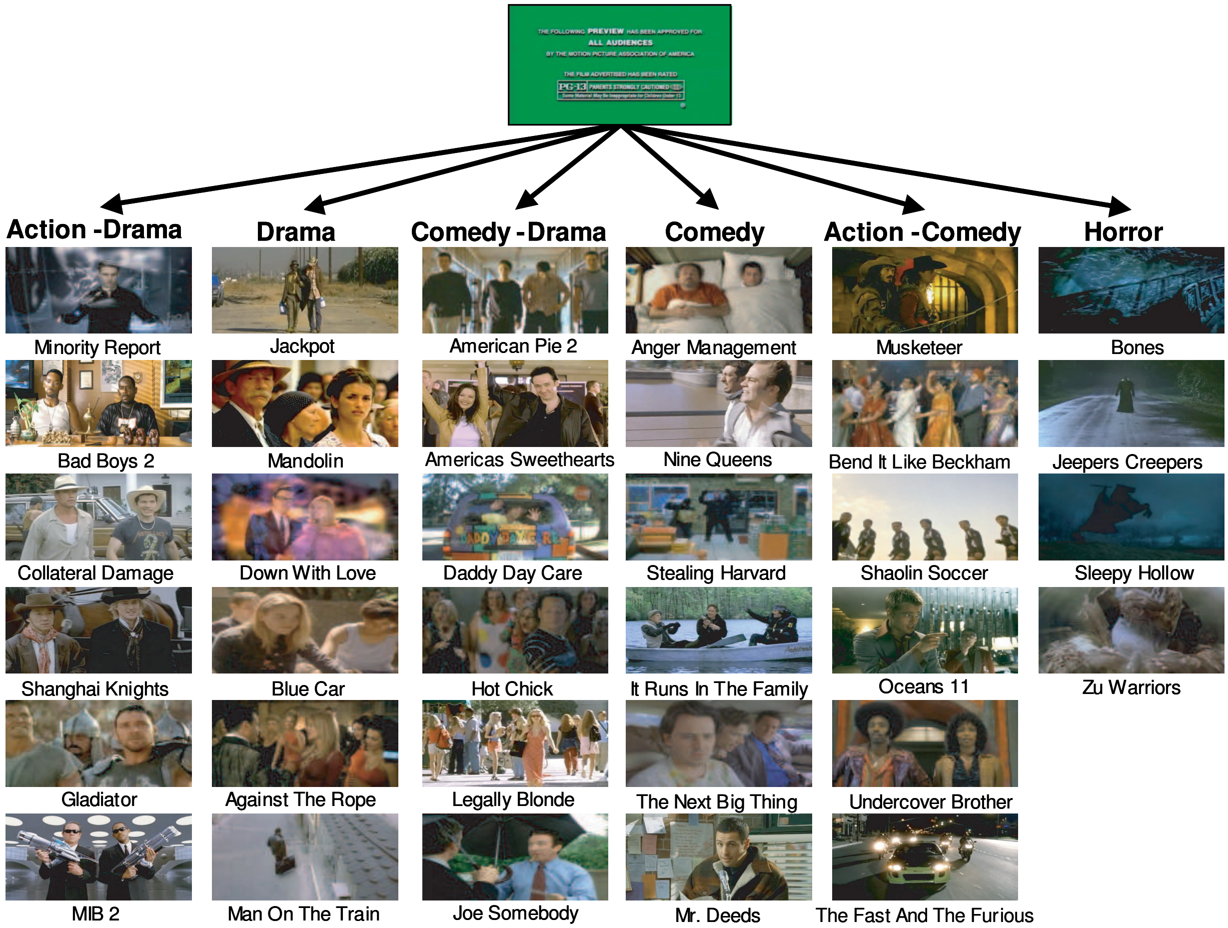
[3] Daniel Arijon. *Grammar of the Film Language*. Hasting House Publishers, NY, 1976.

[4] Internet Movie Data Base. http:// www.imdb.com/.

[5] A. B. Benitez, H. Rising, C. Jrgensen, R. Leonardi, A. Bugatti, K. Hasida, R. Mehrotra, A. Murat Tekalp, A. Ekin, and T. Walker. Semantics of Multimedia in MPEG-7. In *IEEE International Conference on Image Processing*, pages 137–140, vol.1, 2002.

[6] J. S. Boreczky and L. D. Wilcox. A hidden Markov model framework for video segmentation using audio and image features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3741 –3744, 1997.

[7] S. F. Chang, W. Chen, H.J. Horace, H. Sundaram, and D. Zhong. A fully automated content based video search engine supporting spatio-temporal queries. *IEEE Transaction on Circuits and Systems for Video Technology*, pages 602–615, 1998.



(a)

Fig. 11. (a) Examples of film previews in the data set with their associated classes.

[8] Dorin Comaniciu and Peter Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[9] Y. Deng and B. S. Manjunath. Content-based search of video using color, texture and motion. In *IEEE Intl. Conf. on Image Processing*, pages 534–537, 1997.

[10] N. Dimitrova, L. Agnihotri, and G. Wei. Video classification based on HMM using text and faces. In *European Conference on Signal Processing*, 2000.

[11] S. Fischer, R. Lienhart, and W. Effelsberg. Automatic recognition of film genres. In *Third ACM International Multimedia Conference and Exhibition*, pages 367–368, 1995.

[12] Niels Haering. A framework for the design of event detections, (Ph.D. thesis), 1999. School of Computer Science, University of Central Florida.

[13] Informedia. *Informedia Project, Digital video library*. http:// www. informedia. cs.cmu.edu.

[14] B. Jahne. *Spatio-tmporal Image Processing: Theory and Scientific Applications*. Springer Verlag, 1991.

[15] A.K. Jain, R.P.W Duin, and Jianchang Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, Jan 2000.

[16] V. Kobla, D.S. Doermann, and C. Faloutsos. Videotrails: Representing and visualizing structure in video sequences. In *Proceedings of ACM Multimedia Conference*, pages 335–346, 1997.

[17] Cheng Lu, Mark S. Drew, and James Au. Classification of summarized videos using hidden Markov models on compressed chromaticity signatures. In *ACM International Conference on Multimedia*, pages 479–482, 2001.

[18] J. Nam, M. Alghoniemy, and A. H. Tewfik. Audio-visual content based violent scene characterization. In *IEEE International Conference on Image Processing*, pages 353–357, 1998.

[19] Milind R. Naphade and Thomas S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, pages 141–151, 2001.

[20] N. V. Patel and I. K. Sethi. *The Handbook of Multimedia Information Management*. Prentice-Hall/PTR, 1997.

[21] A. Pentland. Finding the illuminant direction. *J. Optical Soc. of Am.*, pages 448–455, July-Sept 1982.

[22] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, July 1990.

[23] R. Qian, N. Haering, and I. Sezan. A computational approach to semantic event detection. In *IEEE Computer Vision and Pattern Recognition*, pages 200–206, 1999.

[24] A. F. Reynertson. *The Work of the Film Director*. Hasting House Publishers, NY, 1970.

[25] Wolf Rilla. *A-Z of movie making, A Studio Book*. The Viking Press, NY, 1970.

[26] B. T. Truong, S. Venkatesh, and C. Dorai. Automatic genre identification for content-based video categorization. In *IEEE International Conference on Pattern Recognition*, pages 230–233, 2000.

[27] N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Transactions on Image Processing*, 9(1):3–19, Jan 2000.

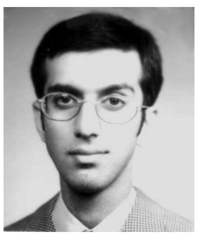
[28] Howhard D. Wactlar. The challanges of continuous capture; contemporaneous analysis, and customzed summarization of video content, 2001. Defining a Motion Imagery Research and Development Program Workshop, sponsored by NCSA/NIMA/NSF.

[29] W. Wolf. Hidden Markov model parsing of video programs. In *International Conference on Acoustics, Speech and Signal Processing*, pages 2609–2611, 1997.

[30] B. L. Yeo and B. Liu. Rapid scene change detection on compressed

video. 5:533–544, IEEE Transaction on Circuits and Systems for Video Technology.
[31] Herbert Zettl. *Sight Sound Motion: Applied Media Aesthetics*. Wadsworth Publishing Company, second edition, 1990.

Zeeshan Rasheed studied Electrical Engineering as an undergraduate at NED University of Engineering and Technology, Pakistan, where he received his Bachelors in Electrical Engineering in 1998. Dr. Rasheed received his Ph.D. degree in computer science from the University of Central Florida, Orlando, in 2003. He was awarded the Hillman Fellowship in 2001 for excellence in research in the Computer Science Ph.D. program. His research interests include video understanding, video categorization, multimedia, human tracking and real time visual surveillance.



Yaser Sheikh received his BS in Electrical Engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan in 2001. He is currently a PhD candidate at the University of Central Florida at the Vision Laboratory. His research interests include georegistration, video segmentation, compression and action recognition. The current emphasis of his work is on correspondence across multiple camera systems.

Mubarak Shah, a professor of Computer Science, and the founding Director of the Computer Visions Lab at University of Central Florida, Orlando, is a researcher in computer vision, video computing and video surveillance and monitoring. He has supervised several Ph.D., MS, and BS students to completion, and is currently directing fifteen Ph.D. and several B.S. students. He has published close to one hundred articles in leading journals and conferences on topics including visual motion, tracking, video registration, edge and contour detection, shape from shading and stereo, activity and gesture recognition and multi-sensor fusion. Dr. Shah is a fellow of IEEE, he was an IEEE Distinguished Visitor speaker for 1997-2000, and is often invited to present seminars, tutorials and invited talks all over the world. He received Harris Corporation Engineering Achievement Award in 1999, the TOKTEN awards from UNDP in 1995, 1997, and 2000; Teaching Incentive Program awards in 1995 and 2003, Research Incentive award in 2003, and IEEE Outstanding Engineering Educator Award in 1997. He is co-author of two books (”Motion-Based Recognition”, Kluwer Academic Publishers, 1997; and ”Video Registration”, Kluwer Academic Publishers, 2003), an editor of international book series on “Video Computing” by Kluwer Academic Publishers, and an associate editor of the Pattern Recognition and Machine Vision and Applications journals. In addition he was an associate editor the IEEE Transactions on PAMI (1998-2002), and a guest editor of the special issue of International Journal of Computer Vision on Video Computing.