Semantic Film Preview Classification Using Low-Level Computable Features

Zeeshan Rasheed Yaser Sheikh Mubarak Shah

Computer Vision Lab
Department of Computer Science
University of Central Florida
Orlando, Fl 32826, USA

Abstract

This paper presents a framework for the classification of feature films into genres, based on computable visual cues. The authors 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 film preview (the 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 Films, Dramas or Horror Films. Computable video features are combined in a framework with cinematic principles to provide a mapping to the four high-level semantic classes. An unsupervised clustering technique is used to discover the structure of the mapping between the computed features and each film genre. Through experiments, we notably demonstrate the structure that exists between low-level features and high-level film genres.

1 Introduction

Directors, actors, and cinematographers use film as a means to communicate a precise 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 it 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. 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"*, [2], 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. *Low-level* symbols in particular like lighting, shot length and background music can be represented in terms of computable features of video data.

Films constitute a large portion of the entertainment industry. Every year about 4500 films are released around the world, which correspond to approximately 9,000 hours of video, [22]. 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 sizable 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 a film containing a few scenes of 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 also 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. In conclusion, we present a framework for genre classification based on computed features from film previews. 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 2. In Section 3, we present the computable video features that are used for classification in our work. Section 4 details use of mean shift classification as a clustering approach in our application. A discussion of the results is presented in Section 5, followed by conclusions in Section 6.

2 Previous Work

One of the earliest research efforts in the area of video categorization and indexing was the Informedia Project [10] 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 [8] and a similar approach by Truong *et al* in [20], 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.

The use of Hidden Markov Models has been very popular in the research community for video categorization and retrieval. Naphade *et al* [13] proposed a probabilistic framework for video indexing and retrieval. Low-level features are mapped to high-level semantics as probabilistic multimedia objects called *multijects*. A Bayesian belief network, called a *multinet*, is developed to perform semantic indexing using Hidden Markov Models. Some other examples that make use of probabilistic approaches are [23, 7, 4]. Qian *et al* also suggested a semantic framework for video indexing and detection of *events*. They presented an example of *hunt* detection in videos, [17].

Specific to film classification, Vasconcelos *et al* proposed a feature-space based approach in [21]. In this work, two features of the previews, average shot length and shot activity, were used. In order to categorize movies they use a linear classifier in the two-dimensional feature space. An extension of their approach was presented in Nam *et al*, [12], 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. The authors 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.

3 Computable 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. We identify four *major* genres, namely Action, Comedy, Horror and Drama. 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.

3.1 Shot detection and Average shot length

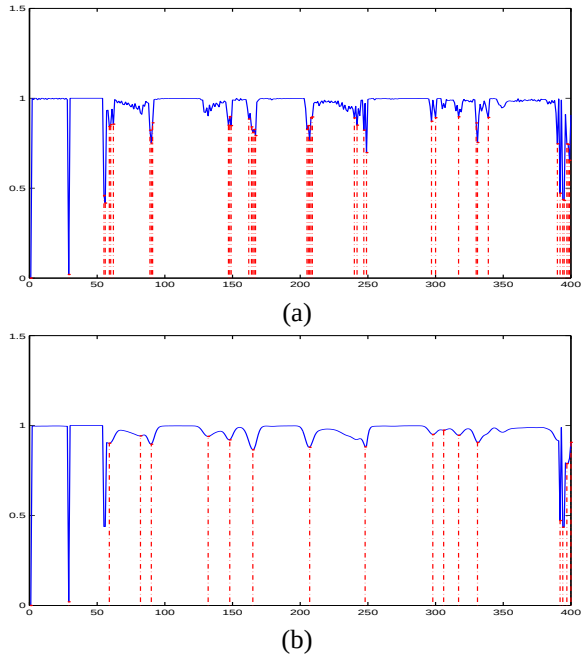
The first feature we employ is the average shot length. This feature was first proposed by Vasconcelos in [21]. The average shot length as a feature represents the tempo of a scene. The director can control the speed at which the audience's attention is directed by varying the tempo of the scene, [20]. 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 [14] for the detection of shot boundaries using

HSV color histogram intersection.

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:

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

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 [14]) 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



(a)

(b)

Figure 1: Shot detection using fixed threshold method for the movie *Red Dragon*. 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 +ve: 25, False -ve: 2 (b) Shots detected by proposed method. Number of shots detected: 18, Correct: 16, False +ve: 2, False -ve: 1

iterative smoothing of the one dimensional function S is performed first. We have adapted the algorithm proposed by Perona *et al* [16] based on anisotropic diffusion. This is done in the context of scale-space. S is 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:

$$\nabla_E S(i) \equiv S(i+1) - S(i),$$
$$\nabla_W S(i) \equiv S(i-1) - S(i). \quad (3)$$

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

$$c_E^t = g\left(\mid \nabla_E S^t(i) \mid \right),$$
$$c_W^t = g\left(\mid \nabla_W S^t(i) \mid \right), \quad (4)$$

where $g(\nabla S) = e^{-(\frac{|\nabla_E|}{k})^2}$. 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 1 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).

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.

3.2 Color Variance

Zettl observes in, [25], '*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 exploiting the variance of color in a clip *as a whole* to discriminate between genres of a film. Intuitively, the

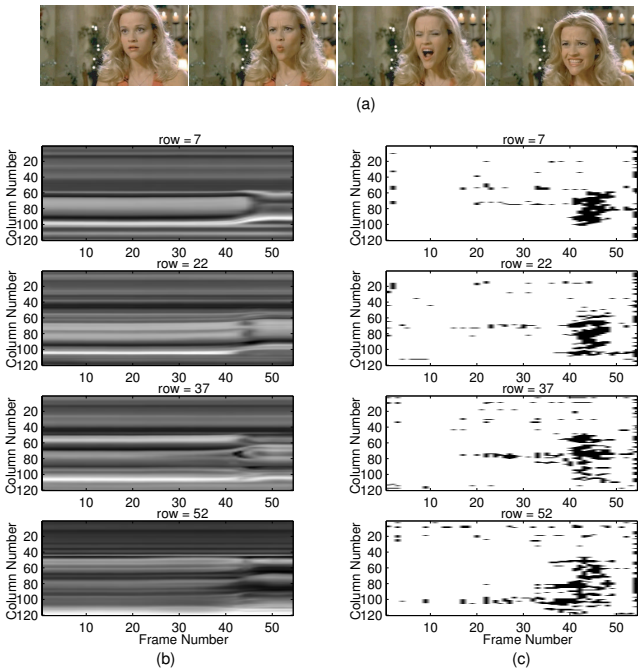


Figure 2: Plot of *Visual disturbance*. (a) Four frames of shots taken from the movie *Legally Blonde*. (b) Horizontal slices for four fixed rows of a shots from the preview. Each column in the horizontal slice is a row of image. (c) Active pixels (black) in corresponding slices.

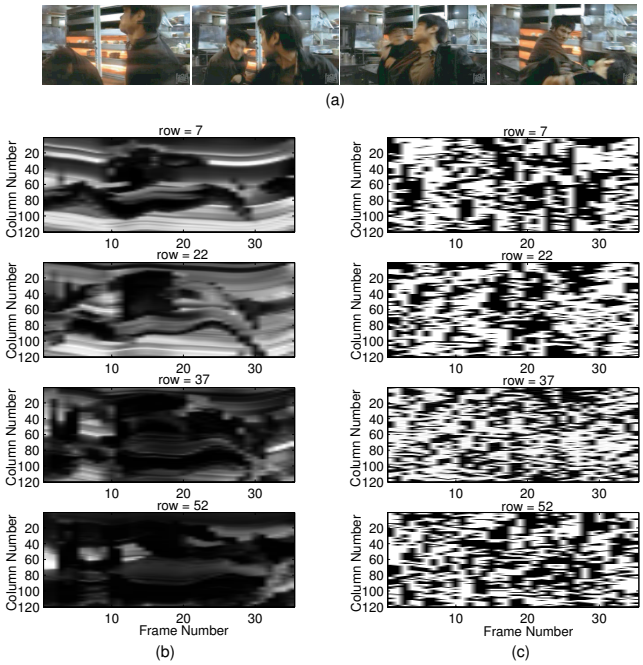


Figure 3: Plot of *Visual disturbance*. (a) Four frames of shots taken from the movie *Kiss of the Dragon*. (b) Horizontal slices for four fixed rows of a shots from the preview. Each column in the horizontal slice is a row of image. (c) Active pixels (black) in corresponding slices.

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 features 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. The covariance matrix of the multi-variate vector 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,

$$\sum = \det(\rho) = \sigma_L^2\sigma_u^2\sigma_v^2 - \sigma_L^2\sigma_{uv}^4 - \sigma_v^2\sigma_{Lu}^4 - \sigma_u^2\sigma_{Lv}^4 + 2\sigma_{Lu}^2\sigma_{Lv}^2\sigma_{uv}^2. \quad (6)$$

This feature is used as a representation of the color variance.

3.3 Motion Content

The *visual disturbance* of a scene can be represented as the motion content present in it. 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 the motion content, an approach based on the structural tensor computation is used, which was introduced in [11]. 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. 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 fall into one bin. On the other hand, pixels with motion other than global motion 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 2 and 3 show motion content measure for two shots. Figure 2 is a dialogue shot taken from the movie *Legally Blonde*. On the other hand, Figure 3 is a shot taken from a fight scene with high activity. Compare the density of moving pixels (black pixels in (c)) of both figures. The density of motion is much smaller for a non-action shot as compared to an action shot.

3.4 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, [25]. A deliberate relationship exists, therefore, between the lighting and the genre of a film.

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. 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.*" [19], p.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, 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.*" [18] p. 96. Many algorithms exist that compute the position of a

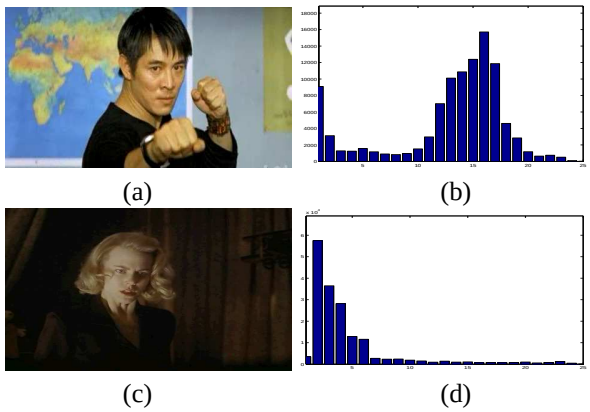


Figure 4: Distribution of gray scale pixel values in (a) *high-key* shot (b) histogram and (c) *low-key* shot (d) histogram.

light source in a given image. 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. However, it is still possible to compute the *key* of lighting using simple computations. 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 4 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,

$$\zeta_i = \mu_i \cdot \sigma_i. \quad (10)$$

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.

4 Mean-Shift Classification

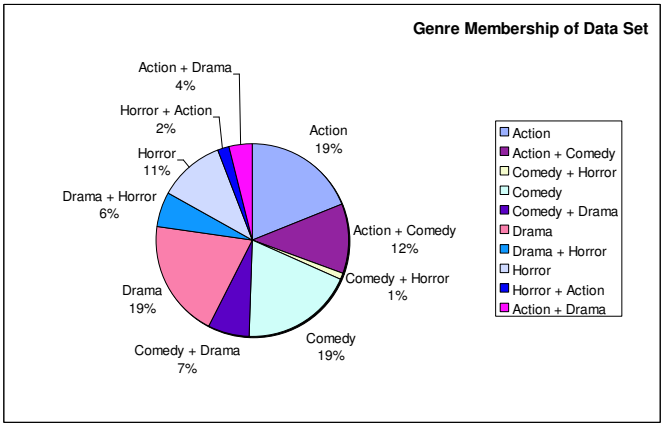


Figure 5: Genre Membership of Data Set. It should be noted that some films have a second genre, as well.

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 the authors 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, 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 [6]. 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. A normal kernel is used, giving a mean shift vector of

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

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*. Such a framework facilitates the study of the mapping between low-level computable features and high-level semantic classes.

5 Results and Discussion

We have conducted extensive experiments on just over a hundred film previews, obtained from the Apple web-site, [1]. 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 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 5. As mentioned earlier, classifying movies into binary 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

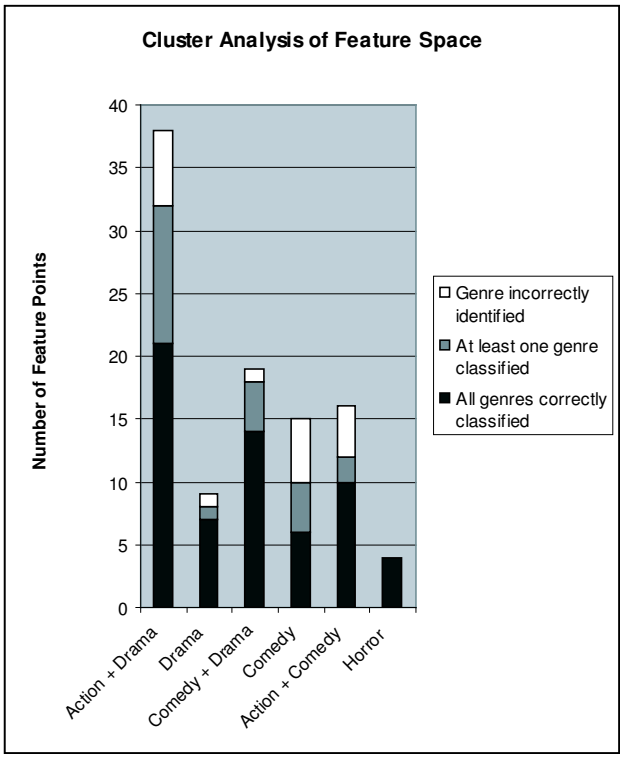


Figure 6: Cluster Analysis of Feature Space. Six clusters are observed in the data and each cluster is classified by its dominating genre.

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 6. 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 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.

The total number of outliers in the final classification was 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.

6 Conclusions

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 film making to discover the higher level description of the entire storielines. 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.

References

[1] http://www.apple.com/trailers/

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

[3] A. B. Benitez, H. Rising, C. Jrgensen, R. Leonardi, A. Bugatti, K. Hasida, R. Mehrotra, A. Murat Tekalp, A. Ekin, T. Walker, *" Semantics of Multimedia in MPEG-7"*, Proceedings of IEEE 2002 Conference on Image Processing (ICIP-2002), Rochester, New York, USA, Sep 22-25, 2002.

[4] J. S. Boreczky and L. D. Wilcox. *"A hidden Markov model framework for video segmentation using audio and image features"*, In Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'97), Seattle, 1997.

[5] S-F. Chang, W Chen, H. J. Meng, H. Sundaram, and D. Zhong, *"A fully automated content based video search engine supporting spatio-temporal queries"*, IEEE Trans. CSVT, vol. 8, no. 5, pp. 602–615, 1998.

[6] D. Comaniciu, P. Meer, *"Mean shift: a robust approach toward feature space analysis"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603-619, 2002.

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

[8] W. Effelsberg, S. Fischer, and R. Lienhart. *"Automatic recognition of film genres"*. In the Third ACM International Multimedia Conference and Exhibition (Multimedia 1995), pages 367-368, New York, Nov. 1995. ACM press.

[9] http://www.imdb.com/

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

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

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

[13] Milind R. Naphade, Thomas S. Huang, *" A probabilistic framework for semantic video indexing, filtering, and retrieval*, IEEE Transactions on Multimedia 3(1): 141-151 (2001)

[14] Niels Hearing, *"A Framework for the Design of Event Detections"* Ph.D. Thesis, School of Computer Science, University of Central Florida, 1999.

[15] N. V. Patel and I. K. Sethi, *"Video segmentation for video data management"*, in The Handbook of Multimedia Information Management, W. I. Grosky, R. Jain, and R. Mehrotra, Eds. Upper Saddle River, NJ: Prentice- Hall/PTR, 1997, pp. 139165.

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

[17] R. Qian, N. Hearing, and I. Sezan, *"A computational approach to semantic event detection"*, in Proc. Computer Vision and Pattern Recognition, vol. 1, Fort Collins, CO, June 1999, pp. 200206.

[18] Wolf Rilla, *"A-Z of movie making"*, A Studio Book, The Viking Press, New York, 1970.

[19] A. F. Reynertson, *"The Work of the film director"*, First Edition. 1970, Hastings House.

[20] B. T. Truong, S. Venkatesh, and C. Dorai, *"Automatic Genre Identification for Content-based video Categorization"*, ICPR 2000.

[21] Vasconcelos, N.; Lippman, A. *" Statistical models of video structure for content analysis and characterization"*,Image Processing, IEEE Transactions on , Volume: 9 Issue: 1 , Jan. 2000 Page(s): 3-19.

[22] Howhard D. Wactlar, *"The Challanges of Continuous Capture, Contemporaneous Analysis, and Customzed Summarization of Video Content"*, Computer Science Department, Carnegie Mellon Univeristy, Pittusburgh, PA 15213, USA.

[23] W. Wolf, *"Hidden Markov Model Parsing of Video Programs"*, Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Munich, Germany, April 1997, pp. 2609-2611.

[24] B. L. Yeo and B. Liu, *"Rapid scene change detection on compressed video"*, IEEE Trans. Circuits Syst. Video Technol., vol. 5, pp. 533 544, Dec. 1995.

[25] Herbert Zettl, *"Sight Sound Motion, Applied Media Aesthetics"*, Second Edition, 1990.