

# CINEMATIC-BASED MODEL FOR SCENE BOUNDARY DETECTION

JIHUA WANG, TAT-SENG CHUA AND LIPING CHEN

*School of Computing  
National University of Singapore, Singapore 117543  
{wangjihu, chuats, chenlp}@comp.nus.edu.sg*

Most current video retrieval systems use shot as the basic unit for information organization and access. A shot, however, models only a visually contiguous sequence of video frames with no coherent semantic meanings. On the other hand, viewers tend to “view” video contents in terms of episodes or scenes. In cinematography, scene is the basic story unit that the directors use to compose and convey their ideas. This paper proposes a cinematic-based model to analyze video contents and extract scene boundaries. Using the list of shots as the base, we first locate rough scene boundaries based on visual similarity. Next, we employ the cinematic rules computationally to refine and cluster the rough scenes. We use a combination of camera focal length and similar shot sequence information to derive the final set of scenes. We test our system on two videos of about 44 minutes in duration. The system has been found to be effective.

## 1 Introduction

The availability of low-cost digital video recording devices has made it possible for most people to enjoy recording, storing, and sharing video information. While the popularity of world-wide-web promotes the proliferation of digital libraries, the combination of these two developments has resulted in the production of huge quantity of on-line information, especially video. The utilities to access video, however, lag far behind the technologies for its creation and delivery. Access to video is still essentially based on shot, which does not match the viewers’ mental model of video contents. There is thus a strong need to develop effective techniques to analyze the contents of video to extract semantically meaningful units to facilitate users’ ad hoc navigation.

Video, like motion picture, derives its meaning from the proper temporal sequencing of frames. “Motion picture communication is discursive. It is unlike a piece of sculpture or a painting, because it achieves its effects by series of images shown over a period of time, and the significance or meaning of any one shot depends both upon what precedes it and what follows it. The principle was applied to motion pictures soon after they were invented...” (Mercer 1971).

Most current video retrieval systems use shots as the basis to organize video contents (Yeung & Liu, 1995 and Zhong et al, 1996). Although shots define contiguous visual units for content structuring, they do not convey coherent semantic contents that match the viewers’ cognitive model. Viewers see and remember video

not in a shot by shot manner, but in terms of events and episodes. In particular, episodes provide natural semantic segmentation of video. Here, we use the term “scene” prevailing in cinematography to denote episode. In a way, shots segments video contents syntactically in terms of visual contiguous units, while scenes model video contents in terms of semantic units that the viewers can associate with. In other words, shots organize video contents at the syntactic level, while scenes target at semantic and viewers level.

Scene is “usually composed of a small number of interrelated shots that are unified by location or dramatic incident” (Beaver 1994). In order to convey idea that has strong resonance with the viewers, Montage is widely used as the basis to model scenes. Montage is developed during the emergence of motion pictures at the beginning of last century when the audio track is not available. According to Mercer’s book pertaining to cinematography (Mercer 1971), “the term Montage has at least two meanings: In America, it refers to a sequence of shots, often with special effects, which communicate to condensed form the general idea that something is taking place. . . . In the European sense, Montage means simply the way shots are put together. It also refers to the physical act of selecting and splicing shots. . . .”. Montage therefore refers to a model that defines the usage of editing effects and camera motions, and the combination and juxtaposition of shots, to evoke the consensus and feeling of spectators and audiences. “Montage is a mighty aid in the resolution of the task of presenting not only a narrative that is logically connected, but one that contains a maximum of emotion and stimulating power.” (Eisenstein 1968). In the complex shots combination, Montage helps film directors express their ideas to the audiences and stimulate them for further understanding and imagination.

In most situations, Montage can be simplified as a set of cinematic rules. This is particularly true in documentary or live sports videos that tend to employ a simple set of rules to construct the scenes. Commonly used rules include (Chua & Ruan 1995): (a) Parallel rule: that aims to convey multiple related activities simultaneously like the chasing or hunting scenes. (b) Concentration or enlargement rule: that presents the context before zooming into the details of the main subjects, and vice versa. (c) Content rule: that models scenes taking place at the same time and location. The main objective of this research is to use Montage as the basis to emulate the creative process of human Directors in composing video. In particular, we plan to divide a long video sequence into shots, and use the set of cinematic rules to analyze the video contents to “uncover” the scenes conceived by the Directors to convey the semantics of video. By using the cinematically modeled scenes as the basic units to organize video, we believe that we are moving a step closer to facilitating effective browsing and retrieval of video by general users. The main contribution of our work is in developing a computationally procedure that uses cinematic models to analyze video contents and extract scene boundaries.

The rest of this paper is organized as follows. Section 2 reviews related work in scene segmentation. Section 3 examines the use of cinematic rules derived from the

theory of Montage to model human director's creative process. Section 4 presents the computational model to perform video content analysis and scene boundary detection. The results of experiment are discussed in Section 5. Finally, Section 6 concludes the paper.

## 2 Related Works

Research on video builds on our knowledge of structural organization of video in terms of shots and scenes, etc. As mentioned previously, shot is a fundamental unit in video capturing, editing and organization. Shots, however, are syntactic units derived purely based on visual continuity. They generally do not have coherent semantic meaning.

In order to derive higher-level semantic entities, a number of recent works investigated the extraction of scenes. As techniques to segment video sequence into shots are well developed (Chua et al 2000), most early researches used shots as the basis to construct scenes. In general, scene boundary detection techniques can be broadly classified into two categories: clustering and segmentation. Most of the existing techniques belong to the clustering category (Yeung & Liu 1996, Zhong et al 1996, Rui et al 1998, Hanjalic et al 1999). These techniques make use of the internal homogeneity of a scene to cluster similar shots together based on visual similarity and time locality. Techniques under the segmentation category examine the external heterogeneities between different scenes. One such technique (Kender & Yeo 1998) proposed a method to calculate shot coherence and use local minima in this continuous measure to detect scene boundaries. The common idea among these techniques is in grouping shots sharing common visual features into scenes. For efficiency reasons, the similarities between shots are computed on the basis of similarity between key-frames or selected key-frames. In order to overcome the limitation of key frame-based matching, Chen & Chua (2001a) modeled the contents of shots as trajectories of content features and developed an efficient string-matching algorithm to identify similar shots. They employed pairs of tiling windows to locate scene boundaries that exhibit the greatest dissimilarity between shots in adjacent tiling windows. The resulting technique is very general and effective. These techniques are able to handle only simple scenes containing shots that share high degree of visual similarity. They are unable to handle the more complex scenes that involve the zooming and panning of contents (concentration/enlargement rule), or featuring multiple simultaneous themes (parallel rule).

In order to overcome the above restrictions, several approaches incorporated techniques to model parallel scenes frequently used in documentaries to present multiple related activities. Rui et al (1998) first clustered visually similar shots into groups, which might not be contiguous. They then merged overlapping groups into scenes to capture parallel scenes involving conversation etc. Hanjalic et al (1999) used the idea of linking similar shots together into treads, and created scenes that

consisted of shots coming mostly from one or more interleaving threads. Yeung et al (1996) employed the idea of Montage, while Yoshitaka et al (1997) considered the grammar of film explicitly to construct scenes consist of similar shots, or alternation between two kinds of shots. These techniques extended the existing work to handle scenes constructed using general content rule or parallel rule. They are tested only on short clips of selected videos rather than the original full-length videos. The techniques to discover parallel scenes tend to be specific to certain types of parallel scenes such as the conversation or chasing scenes. They also have no notion of time locality or used empirical threshold to express time locality, which makes it doubtful whether these techniques are sufficiently general to handle full-length videos. Moreover, they are unable to handle other types of cinematic rules such as concentration/ enlargement rules.

### **3 Cinematic Model for Video Scene Composition**

This section provides an overview of the use of theory of Montage for scene composition, in order to provide a basis for the design of a cinematic-rule-based model for detecting scene boundaries.

Montage emerged in the era when motions pictures were shown without sound. It relies heavily on a carefully structured linear order of shots to create a narrative film sequence to tell a story. Montage, among others, aims to create the following sequencing effects:

- Changes in time: It uses the duration of shots to create the effects of passing of time such as fast, slow, reminiscence.
- Changes in space: It uses shots of varying camera focal distances and angles to establish the spatial relationship between the subjects and the environment.
- Rhythm: It juxtaposes shots with different durations, often involving multiple themes, to create the corresponding slow or quick rhythm.
- Ideology: It uses the paralleling shots from similar or contrasting themes to create the association of ideas and the strong relation between the two subjects.

In other words, Montage theory helps in assembling the shots into a smooth sequence in physical time/space, and in the association of ideas psychologically.

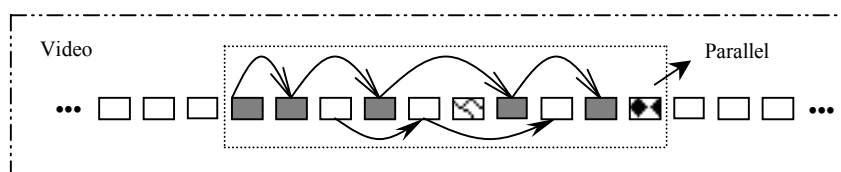
A typical scene involves an activity or subject, together with its context or environment. Two major types of components in a scene are therefore:

- Context: This refers to the information related to the environment of a scene. It includes: (i) date/temporal information; (ii) location which tells the place where the “scene” or story happens; and (iii) other environment information.
- Protagonist: Protagonists are the main objects of the scene. They might be people, animals or other objects that are the main focus of the scene, such as the caribou in the wild life video, or the “father and son” in a family scene.

Although the automatic identification of different kinds of protagonists is not possible, we can use the concept of shot similarities to infer the appearance of same protagonist in similar background.

From the content components and the theory on Montage, we can identify a set of frequently used cinematic rules. The set of rules often studied are (Davenport et al 1991, Chua & Ruan 1995):

- Parallel rule: It is used to compose scenes involving multiple themes, where shots from different themes are shown alternately within the same linear sequence. It provides a powerful tool to show strong relationships between the two subjects. The rule is frequently used to model, say, conversation between two parties, or the hunting scenes in documentary. In particular, the chasing scene involves the alternate showing of the victim and pursuer shots over and over to depict the progression of the chase. The shots belonging to each theme tend to have strong visual similarity and of the same focal distance. Thus such scene can be detected by the repeated showing of two different types of shots within a close proximity as shown in Figure 1.



**Figure 1.** Parallel rule, interaction of two protagonists

The durations of the shots play an important role in determining the rhythm of a parallel scene. Successive shots in short durations are used to model fast tempo, and a closer and more intense relationships between the two protagonists. On the other hand, long shot durations depict a slower leisurely pace in the interplay between the two themes.

- Concentration rule: It always starts with long distance shot, and progressively zooms into close up shots of the main objects. It is used to introduce the main objects and their context. In this process, the camera focal distance of the shots is becoming smaller.
- Enlargement rule: It is the reverse of the concentration rule. It is also used to show the main object and its environment by progressively zooming out from the close up shots of main object. It is used to introduce the context of the current main object before switching to other objects, possibly sharing similar context. Thus, it typically signals the transition to a new scene. During the enlargement, the focal distance of the shots increase progressively.
- General rule: It is the combination of Concentration follow by Enlargement rules. It intends to present an intact action in a sequence. It is normally used to present an event and quickly switch to a new one.

- Serial content rule: This is the most common type of rules used to model scenes that preserve the continuity of location, time, space, and topic. Generally, it shows what goes on in a simple event. Such scene may consist of only a few shots sharing high visual similarity and continuity.

Together, these rules can be used to model most types of scenes appear in documentary, sports and TV serials. They can therefore be used as the basis to “discover” most such scenes in video.

#### **4 The Detection of Scene Boundaries**

This Section describes the details of applying the cinematic model to detect scene boundaries. The main ideas here are to divide the video sequence into shots, and perform the shot-based clustering methods to identify scenes. We view shots as basic lexical units in video, analogous to words or phrases in text, and scenes as semantic units, analogous to sentences or paragraphs in text. In text processing, the locality of words are important in determining the meaning of paragraph, and words that are far away are unlikely to have any effects on the meaning of a local paragraph. This locality constraint applies to scene as well and shots far away will have little effects on the semantic of this scene.

We therefore first merge shots within close proximity into scenes with visually similar contents. This approach is able to detect most of the scenes correspond to simple events, defined using the Serial Content rule, that take place in the same location with similar background (Hanjalic et al 1999). However, this approach tends to over-segment those scenes composed using the Concentration/ Enlargement rules. This is because the sequence of shots appearing in a Concentration/ Enlargement scene is quite different visually. Our next step is therefore to apply the cinematic rules, together with knowledge of camera focal length information, to identify scenes defined using the more complex cinematic rules.

Typical full-length videos on TV may contain many commercial blocks. The TV commercials are composed using a very different set of cinematic rules. Fortunately, there are regular patterns to identify the boundaries of commercial blocks. Here we used the method developed in Koh and Chua (2000) that uses a combination of black frames, static frames and audio silence to segment commercial blocks. The method has been found to be reliable in filtering out most commercial blocks.

Our model-based scene boundary detection method consists of the following steps.

- a) We segment the video into shots. Here we employ the multi-resolution analysis method developed in Chua et al (2000) to segment the shots. The method has been found to be effective in locating both abrupt and gradual transition boundaries. The detection threshold of this method can be tuned to over-

segment the shots, which provides a good base for subsequent steps to merge shots/scenes into higher-level scenes.

- b) We filter out the commercials using the method developed in Koh & Chua (2000).
- c) We merge the shots into a set of scenes using the visual similarity criteria. We called the resulting set of scenes, the scene segments. This method is effective in identifying scenes composed using Serial Content rule and Parallel rule.
- d) Finally, we apply the cinematic rules to merge the above list of scene segments to identify boundaries of scenes composed using the Enlargement / Concentration rules.

The remaining of this Section describes the details of steps (c) and (d) in our procedure.

#### *4.1 The Clustering of Shots into Visually Similar Scene Segments*

Given a list of shots, the first task is to identify scenes using visual similarity and time locality criteria. By considering only scenes that take place in the same location with similar background, we can view a scene as a sequence of shots, where most of the shots share some forms of visual continuity. Here we employ a method developed in Chen & Chua (2001a) to compute shot similarities, and Chen & Chua (2001b) cluster the group of similar shots into scenes using the tiling window method. The tiling window helps to enforce the locality constraint.

##### *4.1.1 Shot Similarity Measures*

At the shot level, we expect two similar shots to exhibit both visual and temporal similarity. Most existing video retrieval methods focused on comparing video shots using either representative or key frames (Aoki et al 1996), and incorporating temporal information around key frames (Jain et al 1999, Zhong et al 1996). As video is rich in both spatial and temporal information, these methods lack full temporal information to support effective shot retrieval. Other methods exploited temporal information more explicitly by modeling the video clips directly as a sequence of video frames (Shan & Lee 1998). However, such methods tend to be inefficient and are unable to model partial similarity between shots.

In our approach, we want to support efficient matching of both exact and partially similar shots. Thus, we want to model not only the content of each frame within the shot, but also the temporal variations of contents across the entire shot. In order to achieve both efficiency and effectiveness, we: (a) model the content of each frame using three visual feature values, the 1<sup>st</sup> and 2<sup>nd</sup> color moments, and a Fractal Dimension texture value (Sarkar & Chaudhuri 1994); and (b) model the content of the entire shot as the trajectories of these three quantized feature values. For each feature, we employ the efficient longest common sub-sequence (LCS) matching

algorithm to find the length of LCS between two input trajectories belonging to two different shots. The LCS found at the end of this algorithm measures the number of frames matched between the two shots, while ignoring those not matched. Thus, one good measure of similarity between two shots,  $S_q$  &  $S_d$ , is simply the proportion of frame matches between the two shots as follows:

$$\text{Sim}_i(S_j, S_k) = \frac{\text{LCS}_i(|S_j|, |S_k|)}{\text{Min}(|S_j|, |S_k|)} \quad (1)$$

We apply Equation (1) to all the three feature representations of the frame sequence. By assigning appropriate weight  $w_i$  to the different features, we can derive the overall similarity between the two shots as:

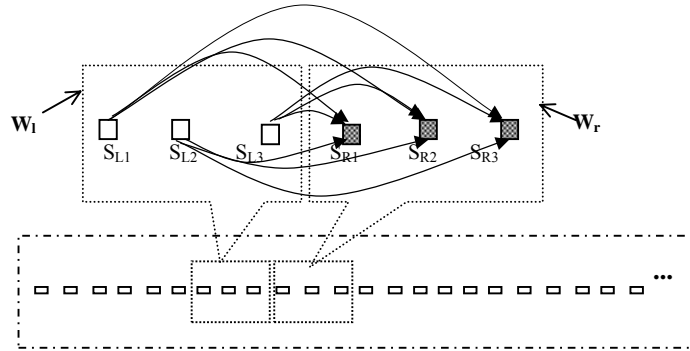
$$\text{Sim}(S_j, S_k) = \sum_{i \in [\text{feature}]} \text{Sim}_i(S_j, S_k) \cdot w_i ; \text{ where } w_i > 0 \text{ and } \sum_{i \in [\text{feature}]} w_i = 1 \quad (2)$$

The algorithm has been found to be effective in locating similar shots of varying lengths.

#### 4.1.2 Sequence Comparison for Scene Segments

A simple video scene consists of a sequence of semantically related shots, unified by visual similarity and time locality (Rui et al 1998). Visual similarity, in terms of both spatial and temporal similarity, is enforced by the shot matching algorithm. Time locality is enforced by employing the sliding window approach. If we consider a pair of sliding windows, one on the left and one on the right, then the boundaries of visually similar scenes occur at positions whereby the contents of the left window are most dissimilar to that on the right (see Figure 2). We can compute the similarity between the contents of left and right windows as follows. Let  $W_L$  and  $W_R$  be the set of shots on the left and right window respectively. We first use Equation (2) to compute the similarities between each pair of shots from each window; that is, compute all  $\text{Sim}(S_i, S_j)$  values, where  $S_i \in W_L$ ,  $S_j \in W_R$ . Next we compute the similarity between  $W_L$  and  $W_R$  as:

$$\text{Sim}(W_L, W_R) = \frac{1}{|W_L| \times |W_R|} \sum_i \sum_j \text{Sim}(S_i, S_j) \quad (3)$$



**Figure 2.** Similarity comparison between two tiling windows

Here the choice of tiling window size needs careful consideration. Our experimental results show that a good choice for window size is 3 shots.

The overall sliding window algorithm to detect scenes, similar to the text tiling method for locating text segments (Hearst & Plaunt 1993), is as follows.

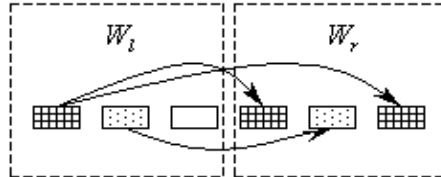
- Move the sliding window pair ( $W_L$ ,  $W_R$ ) over the sequence of shots in the database at one shot increment. At each window position, compute  $\text{Sim}(W_L, W_R)$  using Eqn. (3).
- Plot the sequence of similarity values at each window position, and perform the smoothing using the simple local mean smoothing method with a window size 3 to eliminate small fluctuations in the curve. The resulting curve shows the degree of similarity between left and right sliding windows at each position. The local minims on the curve show possible positions of scene boundaries.
- Use the threshold  $\tau_s$  to remove those local minims where the difference between left and right sliding windows is not significant enough to conclude a scene boundary.
- Select the remaining local minims, at a minimum of  $w_s$  shots apart, as scene boundaries.

The above scene segmentation algorithm assumes that a scene takes place in the same location and share many common backgrounds. Hence the majority of the shots belonging to the same scene possess common visual features, which differ from that of other scenes. This is true in most of the simple scenes composed using the Serial Content rule. However, the algorithm tends to over-segment the more complex scenes composed using, say, the Concentration/Enlargement rules, that involve a lot of zooming and panning where the background of shots change drastically. To handle such scenes, we need to model cinematic rules and consider the camera parameters such as zooming, panning, and focal length etc.

#### 4.2 Detection of Parallel Scenes

Although the algorithm outlined in Section 4.1 is designed to handle only simple visually similar scenes, it is able to locate parallel scenes naturally. A parallel scene contains one or more sequences of inter-leaving shots, depicting multiple

simultaneous events as shown in Figure 3. While we compare the similarity of neighboring windows, the contents of these two windows will be considered as similar as they tend to contain similar repeated set of shots.



**Figure 3.** Tiling windows over a parallel scene

#### 4.3 Detection of Scenes defined using Enlargement/ Concentration/ General Rules

After we have segmented the list of visually similar scene segments, our next task is to locate those scenes defined using Concentration or Enlargement rules. This is accomplished by applying the appropriate cinematic rules to merge the candidate scene segments. In order to apply these rules, we need to know the camera parameters of all the shots in the video sequence. For simplicity, we employ only one camera parameter, the focal distance, for each shot. We estimate the focal distance manually based on the size of main objects. The focal distance ranges from 6 (long distance shot), to 3 (medium distance shot, equivalent to showing the full size of a person), to 1 (close-up shot or equivalent to showing the face of a person on half the screen).

For scenes composed using the Concentration rule, we expect the focal distance of shots to decrease gradually, and it is used to introduce the context where the story happens before showing the main subjects. The scenes defined using Enlargement rule is the reverse, and is typically used to show the context before switching to a new story. The General rule shows the transition from one theme to another. According to the rules, we can analyze the features of scene segments to cluster appropriate segments into one of these complex scenes.

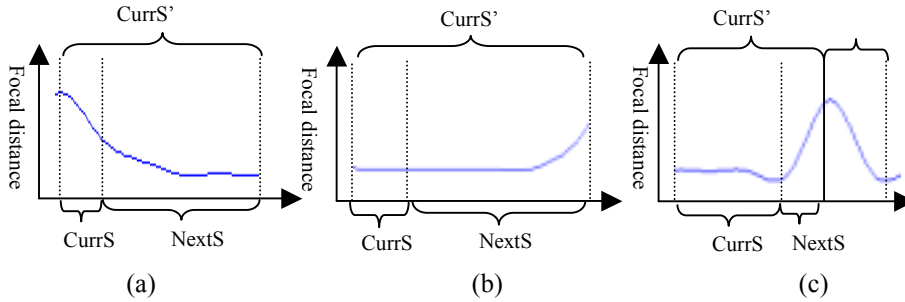
We use *CurrS* to denote the current scene segment under consideration, and *NextS* for the next scene. We initially set *CurrS* to be the first scene segment of the video sequence. The algorithm proceeds as follows.

- a. If the number of shots in *CurrS* is less than a threshold  $\tau_s$ , then: (See Figure 4 for illustration)
  - Case 1 (Concentration rule): If the focal distances of shots reduce steadily from *CurrS* into *NextS*, it indicates that *CurrS* should be merged with *NextS* as part of a Concentration scene.

- Case 2 (Enlargement rule): If the focal distances of shots increase from *CurrS* into *NextS*, then merge the *CurrS* and *NextS* as part of an Enlargement rule.
- Case 3 (General rule): If the focal distances of shots in *CurrS* increase into *NextS*. But the focal distances of the shots in *NextS* exhibiting a peak, showing an increasing follow by decreasing trends, then it indicates that a scene boundary occurs within *NextS*. We divide the *NextS* into two parts separated at the peak (see Figure 4c). We merge the first part with *CurrS* to form a scene, and merge the second part with the following scene segment to form the new *NextS*.

- Proceed to the next scene segment by setting  $CurrS = NextS$ , and assigning *NextS* to the following scene segment.
- Repeat from Step (a) until all the scene segments have been considered.

Figure 4 illustrates the three cases described above. Here, *CurrS'* means the “current scene segment” after merging.



**Figure 4.** Merging of scene segments based on cinematic rules

At the end of the above processes, we obtain a list of scenes satisfying our set of cinematic rules.

## 5 Analysis and Classification of Scenes

We would like to analyze the scene contents and determine its type, based on the cinematic rules used for its composition. To do this, we need to first identify the variety of shots available in the scene, and map the pattern with the appropriate cinematic model. Knowledge of scene structure and its cinematic types will be invaluable to analyzing the semantics of the overall video.

Based on the visually similar criterion defined in Section 4.1, our scenes may contain one or more variety of shots. We first identify the variety of shots within each scene by linking visually similar shots together. The number of links created denotes the variety of shots within the scene. For each  $Link_k$ , we compute its average

link distance,  $AvgDist_k$ . This is done by first averaging the temporal distances  $Dist(S_i, S_j)$  of all pairs of neighboring shots  $S_i, S_j$  within  $Link_k$ . Here  $Dist(S_i, S_j) = |i-j|$ , and  $i, j$  are the original link positions of the shots in the scene.  $Dist(S_i, S_j)$  expresses the distance of each pair of similar shots within the original scene sequence. If most  $Dist(S_i, S_j)$ 's are 1, it means that most shots in the original scene belongs to one shot type. If most  $Dist(S_i, S_j)$ 's are 2, it means that similar shots appear alternately, implying the possibility of parallel scene where two groups of shots appear alternately within the scene.

We can now classify the scenes into one of the following 5 classes.

- Parallel scene: If there are at least two linked sequences of shots, and the average link distance in each link is around 2.
- Concentration scene: If there is only one or no significant link in the scene, and, the focal distances of the shots reduce towards the end of the scene.
- Enlargement scene: It is similar to that of Concentration rule except that the focal distances of the shots increase towards the end of the scene.
- General scene: A combination of the above two cases.
- Serial scene: If there is no significant link, and no Concentration or Enlargement characteristic is observed.

## 6 Experiment Results and Evaluation

We use one full-length movie and part of a documentary to test our proposed scene analysis method. The movies are obtained from the Media Corporation of Singapore (MediaCorp). The movie is an hour-long detective TV series including commercials. The movie is more artistically composed and contains many changes in locations where the story takes place. For both videos, we can observe that clear cinematic rules are used to compose most of the scenes. In order to remove the noise introduced by the commercials, which use different styles of composing the contents, we filter out the commercials before applying our scene detection algorithm.

Table 1: Statistics of Test Videos

	Frame #	Shot #	Scene #	Duration
Video 1: Movie	62,209	521	42	41.5min
Vide 2: Documentary	4322	26	4	2.9min
Overall:	66,531	537	46	44.4 min

In order to test our system objectively, we need to extract ground truth on the boundaries of the scenes. As the scene is a high-level concept, subjectivity might be

involved when the viewers do the segmentation manually. To avoid such problems, we used two human viewers to view the movie independently, and propose scene boundaries. The scenes segmented by the viewers are mostly the same. The differences are resolved through full discussions. We then use the final set of boundaries as ground truth. Table 1 summarizes the statistics of the two test videos. There are altogether 46 scenes in over 44 minutes of video.

Table 2 shows the results of scene boundary detection at the end of two stages. The two stages are: (A) after the clustering of visually similar scenes described in Section 4.1; and (B) after the application of cinematic rules to identify more complex scenes as described in Section 4.3. From Table 2, we can see that at the end of Stage A, we could achieve a high recall of 96%; but the precision is low at 66%. This is to be expected as the tiling window method based on visual similarity criteria tends to over-segment scenes, especially those complex scenes composed using the Concentration and Enlargement rules. After the application of cinematic rules in Stage 2, we are able to improve the precision drastically to 87%, while the recall dips only slightly to 89%. The results clearly demonstrate that the use of cinematic rules is effective.

Table 2. Scenes detected using Scene Segments vs. after Applying Cinematic Rules

	<b>Total</b>	<b>Wrong</b>	<b>Miss</b>	<b>Precision</b>	<b>Recall</b>
Stage A:	67	23	2	66%	96%
Stage B:	47	6	5	87%	89%

Stage A: after cluster visually similar shots

Stage B: after apply cinematic Rules

To illustrate the results, we present and analyze two more complex scenes correctly detected by our method. The first example shows a Parallel rule scene involving the conversation between three persons in the office. There are three links in this parallel scene, two for main objects, and one for the context. There is also a single shot for the third person, but this is not essential (See Figure 5). The three links are inter-leaving, clearly showing the presence of simultaneous actions taking place.

The second example shows a scene of an American Caribou in the wild life documentary. Here the first 4 shots (1 – 4) tell us something about the living condition of the caribou. The focal length zooms in gradually into the closed up shot of a caribou. In the process, audience gains a good the knowledge of the cold and snowy environment where such “animal” lives. The last two shots move from near to far shots to show that the caribous live in the group.

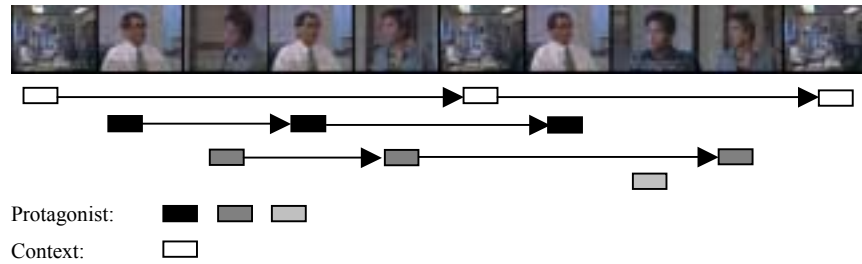


Figure 5. An example of a parallel scene



Figure 6. An example of a general rule scene

Finally, we evaluate the accuracy of the classification method outlined in Section 4.4 to identify the types of scenes generated. Out of the 47 scenes given in the ground truth database, we are able to correctly identify and classify 35 scenes, giving an overall classification accuracy of 74% as listed in Table 3.

Table 3. Scene Classification Result

Total	Correct	Accuracy
47	35	74%

## 7 Discussion and Future Work

In cinematography, scenes are the basic story units that the directors used to compose and convey their ideas. Most viewers tend to “view” video contents in terms of scenes. Thus this research aims to identify scene boundaries and use scenes as the basis to organize video data for intuitive user access. This paper proposes a cinematic-based model to analyze video contents and extract scene boundaries. Using the list of shots as the base, we first locate rough scene boundaries based on visual similarity. Next, we employ the cinematic rules computationally to refine and merge the rough scenes. We use a combination of camera focal length and similar shot sequence information to derive the final set of scenes. We test our system on two videos of about 45 minutes in duration. The system has been found to be effective.

The work reported here represents only the beginning to this line of research. Through our analysis of the scene created, and the books on cinematography, we can see that the camera parameters such as the focal distance and camera angle play an important role in scene construction and discovery. Unfortunately the automatic

recognition of focal length is an unsolved problem and further research on this topic is important. In addition, we also found that the use of film grammar and cinematic rules is essential to characterize the scenes. Our research will also investigate formal models for film grammar, and develop stochastic techniques such as the Hidden Markov Model (Rabiner 1989) to discover scenes in a learning based approach.

## 8 Acknowledgements

The authors would like to acknowledge the support of the National Science and Technology Board, and the Ministry of Education of Singapore for supporting this research under the research grant RP3989903.

## References

1. Arman F.; Depommier, R.; Hsu, A. & Chiu, M-Y.- Content-based Browsing of Video Sequences. *ACM Multimedia* (1994) pp. 97 – 103, Oct.
2. Aoki, H.; Shimotsuji, S. & Hori, O.- A Shot classification Method of Selecting Effective Key-Frames for Video Browsing. *ACM Multimedia* (1996) pp. 1 – 10.
3. Beaver, F.- *Dictionary of Film Terms* (1994). Twayne Publishing, New York.
4. Chen, L.P. & Chuat T.S. (2001a). A Matched tiling Approach to video Retrieval. To appear in *ICME (IEEE Conf. On Multimedia System and Expo)*'2001, Waseda, Japan, Aug. 2001.
5. Chen, L.P. & Chua, T.S. (2001b). Match and Tiling: A Unified Framework for Video Retrieval and Scene Segmentation, *PRIS Report* (2001), School of Computing, National University of Singapore.
6. Chua, T.S.; Kankanhalli, M. & Lin, Y.- A General Frame Work for Video Segmentation Based on Temporal Multi-resolution Analysis, *Int'l Workshop on Advanced Image Technology* (2000) pp. 119 – 124.
7. Chua, T.S. & Ruan, L.Q.- A Video Retrieval and Sequencing System, *ACM Trans. Inf. System*, **13**(4) (1995) pp. 373 – 407.
8. Davenport, G.; Smith, T.A. & Princever, N., Cinematic Primitives for Multimedia, *IEEE Computer Graphics and Applications*, **11**(4), Jul (1991) pp. 67 – 74.
9. Eisenstein, Eergei M.- *The Film Sense*, Faber and Faber Limited, (1968) pp. 14.
10. Hanjalic, A.; Lagendijk, R.L. & Biemond, J.- Automated High-level Movie Segmentation for Advanced Video Retrieval System, *IEEE Trans. on Circuits and System for Video Technology* (1999) pp. 580 – 588.
11. Hearst, M.A. & Plaunt, C.- Subtopic Structuring for Full-Length Document Access. *ACM SIGIR* (1993) pp. 59 – 68.
12. Jain, K.; Vailaya, A. & Wei, X.- Query by Video Clip. *Multimedia Systems* **7** (1999) pp. 369 – 384.

13. Kender, J.R. & Yeo, B.L.- Video Scene Segmentation via Continuous Video Coherence. *Proc. IEEE Int'l Conf. On Computer Vision and Pattern Recognition* (1998) pp. 367 – 373.
14. Koh, C.K. & Chua, T.S.- Detection and Segmentation of Commercials in Video, *UROP Report* (2000), School of Computing, National University of Singapore.
15. Mercer, J.- *An Introduction to Cinematography* (1971), Stipes Publishing Co., pp. 57.
16. Rabiner, L. R.- A Tutorial On Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE*, **77**(2) (1989) pp.257 – 286.
17. Rui, Y.; Huang, T.S. & Mehrotra, S.- Exploring Video Structure Beyond the Shots. *Proc. IEEE Conf. On Multimedia Computing and Systems* (1998) pp. 237 – 240.
18. Sarkar, N. & Chaudhuri, B.B.- An Efficient Differential Box-Counting Approach to Compute Fractal Dimension of Image. *IEEE Trans. on Systems, Man and Cybernetics*, **24** (1) Jan. (1994) pp. 115 – 120
19. Shan, M.K. & Lee, S.Y.- Content-based Video Retrieval Based on Similarity of Frame Sequence. *IEEE Proc. of Int'l Workshop on Multimedia Database Management Systems* (1998) pp. 90 – 97.
20. Yeung, M. & Liu, B.- Efficient Matching and Clustering of Video Shots, *Proc. IEEE ICIP' 95* **1** (1995) pp. 338 – 341.
21. Yeung, M.; Yeo, B.L. & Liu, B.- Extracting Story Units from Long Programs for Video Browsing and Navigation, *IEEE Proc. of Multimedia'96* (1996) pp. 296 – 305.
22. Yeung, M.; Yeo, B. L.; Wolf, W. & Liu, B.- Video Browsing using Clustering and Scene Transitions on compressed Sequences, *Proc. IS&T/SPIE Multimedia Computing and Networking* (1995) pp. 399 – 413.
23. Yoshitaka, A.; Ishii, T.; Hirakawa, M. & Ichikawa, T.- Content-based Retrieval of Video Data by the Grammar of Film, *Proc. IEEE Symposium on Visual Languages* (1997) pp. 310 – 317.
24. Zhong, D.; Zhang H. & Chang, S.F.- Clustering Methods for Video Browsing and Annotation, *Proc. IS&T/SPIE Storage and Retrieval for Still Image and Video Database IV* **2670** (1996) pp. 239 – 246.