The Visual Computer

A cinematic-based framework for scene boundary detection in video

Jihua Wang,
Tat-Seng Chua

School of Computing, National University of Singapore, Singapore 117543
E-mail: {wangjihu, chuats}@comp.nus.edu.sg

Published online: 14 February 2003

Most current video retrieval systems use shots as the basis for information organization and access. In cinematography, scene is the basic story unit that the directors use to compose and convey their ideas. This paper proposes a framework based on the concept of continuity to analyze video contents and extract scene boundaries. Starting from a set of shots, the framework successively applies the concept of visual, position, camera focal distance, motion, audio and semantic continuity to group shots that exhibit some form of continuity into scenes. The framework helps to explain the principles and the heuristics behind most cinematic rules. The idea is tested using the first three levels of continuity to extract the scenes defined using the most common cinematic rules. The method has been found to be effective.

Key words: Cinematic model – Scene detection – Video retrieval

1 Introduction

The availability of low-cost digital video-recording devices and networks have made it possible for most people to enjoy recording, storing, and sharing video information. While the popularity of the World Wide Web promotes the proliferation of digital libraries, the combination of these two developments has resulted in the production of a 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 the 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 and Liu 1995; 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, episodes, and stories. In particular, episodes provide natural semantic segmentation of video. Here, we use the term scene, prevailing in cinematography, to denote an episode. In a way, shots segment 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 viewers at a semantic level.

A 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 an idea that has a strong resonance with the viewers, some cinematic rules and montage are widely used as the basis to model scenes and to keep the content consistent. All these rules, including montage, were developed during the emergence of motion pictures at the beginning of last century when the audio track

The Visual Computer (2003) 19:329–341
Digital Object Identifier (DOI) 10.1007/s00371-002-0184-9

was not available. According to Mercer's book pertaining to cinematography, "the term Montage has at least two meanings: In America, it refers to a sequence of shots, often with special effects, which communicate in 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..." (Mercer 1971). 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 complex shot combination, montage helps film directors express their ideas to the audiences and stimulates 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 the following (Chua and Ruan 1995):

- The parallel rule aims to convey multiple related activities simultaneously, like chase or hunting scenes.
- The concentration or enlargement rule presents the context before zooming into the details of the main subjects, or vice versa.
- The content rule models scenes taking place at the same time and location.

Furthermore, other cinematic "continuity" rules are also well developed and widely used, like position continuity and motion continuity. "Continuity is (used in) maintaining the established flow of story of visual and aural production detail between takes, shots, and scenes" (Thompson 1998). Position continuity serves to preserve the relative positions of the objects within the screen during a scene. Motion continuity ensures that the direction of movement of the dominant objects within the scene is maintained. In fact, the three cinematic rules given earlier can also be viewed as maintaining some form of continuity. All the types of continuity are very critical and cannot be violated unless the story unit is changed.

The main objective of this research is to use montage as the basis to emulate the creative process of the human directors in composing video. From our analysis, it can be seen that continuity is a unifying concept that is used in all the cinematic rules in defining coherent scenes. This research therefore explores a framework based on continuity that can be used to model most cinematic rules. In particular, we plan to divide a long video sequence into shots, and use the concept of continuity in conjunction with the cinematic rules to analyze the video contents to "uncover" the scenes conceived by the directors. 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 the general users. The main contribution of our work is in developing a framework and computational procedures based on cinematic models to 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 including the 180° rule and those derived from the theory of montage to model the human director's creative process. Section 4 presents our overall framework based on the concept of content continuity. Section 5 presents the computational procedures to perform video content analysis and scene boundary detection. The results of experiment are discussed in Sect. 6. Finally, Sect. 7 concludes the paper.

2 Related works

Research on video builds on our knowledge of structural organization of video in terms of shots and scenes. As mentioned previously, the shot is a fundamental unit in video capturing, editing and organization. Shots, however, are syntactic units whose derivation is purely based on visual similarity criteria. 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, the techniques of scene boundary detection can be broadly classified into two categories: clustering and segmentation. Most of the existing techniques belong to the clustering category (Yeung and 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. The criteria used are typically based on the visual similarity arising from the application of 180° rule in film capturing (Hari and Chang 2000) and time locality. Techniques under the segmentation category examine the external heterogeneities between different scenes. One such technique proposed a method to calculate shot coherence and use local minimums in this continuous measure to detect scene boundaries (Kender and Yeo 1998). 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 matching, Chen and Chua (2001) 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. But, overall, these techniques are able to handle only simple scenes containing shots that share high degree of visual similarity.

In order to extend the 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 threads, 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 consisting of similar shots or an alternation between two kinds of shots.

These techniques extended the existing work to handle scenes constructed using the general content and parallel rules. They were 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 chase 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 or enlargement rules.

3 Cinematic model for video scene composition

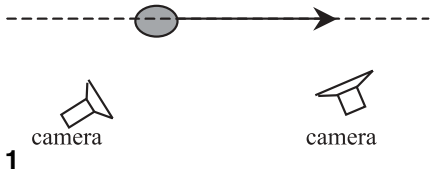
As we mentioned above, a scene consists of one or more shots, and it is used to tell the audience a story in a coherent way. Such scenes in a movie can be regarded as some sub-story that happened with the objects' sequential interactions. Movie directors and editors follow some important rules to compose the scenes to convey semantics of movies consistently. The collection of those cinematic rules is also called film grammar, and these film grammars are widely used to simulate the changes in time, space, topics, and the relation of the objects.

This section provides an overview of the movie composition syntax and cinematic rules in capturing and editing, including the rules for maintaining the content continuity including the 180° rule (Hari and Chang 2000) and the set of montage rules.

3.1 The 180° *rule*

The 180° rule, also called the triangle principle (Arijon 1976), states that the viewpoint should always be on one side of the line between main subjects. In other words, if there are subjects interacting in a shot or the camera is moving, a certain line will be formed. That line determines from where we "look at" the subjects or view the unfolding of the stories. The rule states that the camera should stay to one side of the line thus maintaining the relative positions between the main subjects. Figure 1 illustrates the 180° rule for both static and moving objects.

The advantage of this rule is that the relative position of the subjects on the screen never changes. Thus, the viewer is able to better perceive consistent relationships between the main subjects in time and space. Figure 2 gives an example of a scene where this rule is enforced. We can see clearly that the two persons preserve their relative positions in the talking scene in which the woman looks to the left and the man looks to the right. What would happen if the rule is violated? Figure 3 shows such an example in which viewer may be confused with the flow of the story.



1


2


3

Fig. 1. 180° rule

Fig. 2. The original video shows a good example (Long Vacation, Japanese TV series)

Fig. 3. Negative example composed by us to illustrate the point

The rule has also been applied in computer animations to automate camera control in scenes containing virtual actors (Noma and Okada 1992).

3.2 Montage rules

The theory of montage plays an important role in composing individual shots together to form a meaningful whole. Montage refers to the set of rules used for the editing and composition of shots to convey the intent of the directors. Montage emerged in the era when motion pictures were shown without soundtracks. It relied heavily on a carefully structured linear order of shots to create a narrative film sequence to tell a story. While others may argue that "the advent of sound heralded the death of montage" (Davenport et al. 1991), montage still plays a key role in most video composition nowadays. Montage, among others, aims to create the following sequencing effects:

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

In other words, montage theory helps in assembling the shots into a smooth sequence in physical time and space, and in the psychological association of ideas. 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 or temporal information, (ii) location that 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 wildlife 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 as follows (Davenport et al. 1991; Chua and Ruan 1995):

- *Parallel rule*. The parallel rule 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 documentaries.

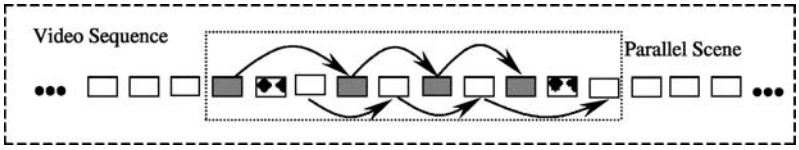


Fig. 4. Parallel rule – interaction of two protagonists

In particular, the chase 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 Fig. 4.

- *Concentration rule*. The concentration rule starts with the 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's focal distance of the shots is becoming smaller.
- *Enlargement rule*. This is the reverse of the concentration rule. It is also used to show the main objects and the 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*. This is the combination of the concentration rule follow by the enlargement rule. 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 a scene may consist of only a few shots sharing high visual similarity and continuity.

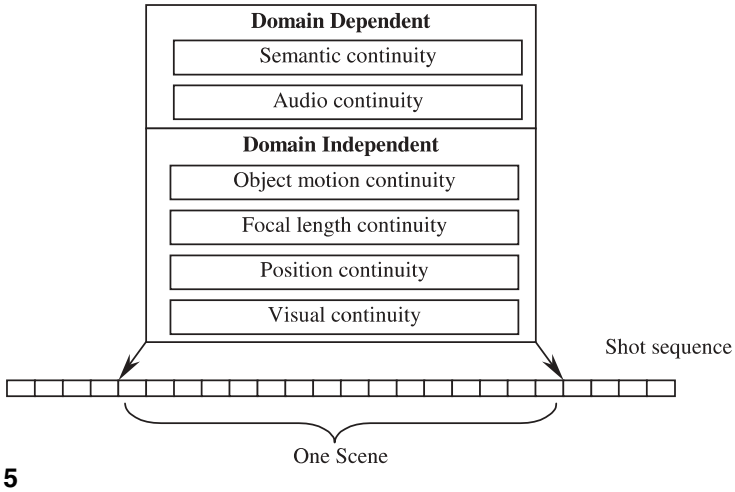
Together, these rules can be used to model most types of scenes that appear in documentaries, sports coverage and TV serials. They can therefore be used as the basis to "discover" most such scenes in video.

4 The overall framework for scene detection

The theory of montage and the cinematic rules have been used as the basis by which directors and editors to put shots together to create coherent stories. From our analysis, it can be seen that continuity is an invariant theme that unifies all cinematic rules. Continuity is manifested in different forms under different rules and to different degree of sophistication. For example, visual continuity is maintained in constructing scenes that take place in the same time and location, while some forms of increasing or decreasing camera focal length continuity is applied in defining concentration or enlargement rules. In addition, view, motion or even audio continuities are also used to construct scenes based on some general or specific rules.

To capture the concept of continuity and to use it to unify most cinematic rules, we develop a framework for scene boundary detection. The framework aims to provide better understanding of the principles behind the cinematic rules and to facilitate the development of computational procedures to extract scenes constructed based on these rules. Figure 5 shows the continuity framework. The relations between each level of continuity and the cinematic rules are explained below:

- *Visual continuity*. Visual continuity exists between successive shots with similar background. Such similar background models the scenes that happen at the same time and location. It also captures the interweaving shots with different objects that hold similar background.
- *Position continuity*. The 180° rule limits the changes in camera angle in filming a scene. It ensures that objects preserve their relative positions in the scene, as any sudden change of the relative position confuses the audience.



5

6

Fig. 5. Content continuity
Fig. 6. Motion continuity is preserved in the four consecutive shots (Beijing Bicycle)

- *Focal distance continuity*. Focal distance typically changes in a continuous way in the establishing stage, as well as the ending stage of a scene. Such continuous change is captured by the simple cinematic rules like the concentration rule at the beginning of a scene, and the enlargement rule toward the end. This kind of continuity is used quite a lot to produce stories in documentary in a relax way. In the movies, especially action movies, other types of shots that introduce the environment tend to be added into the focal distance continuity scenes. It tends to make the detection of such scenes in action movies more complex.
- *Object motion continuity*. Object motion continuity has a strong relation to the 180° rule. Here the direction of motion of the main object within a scene should be the same. Figure 6 shows a two-person chase scene in the four consecutive shots in which the two men both run from left to the right.
- *Audio continuity*. The sound track in a movie contains environment sound and dialogue. The same scene should possess the similar environment sound and dialogues by the same speakers, especially for dialogue scenes. This rule might not be true for other types of scene such as those in documentaries where the same dialogues tend to be carried over to the new scene.
- *Semantic continuity*. At the highest level, scenes are constructed based on semantic coherence, such as news video, etc. Semantic continuity is a high-level concept and is dependent on the content of video. Domain knowledge is therefore needed to uncover the scenes based on semantic continuity.

The framework is designed in such a way that the lower-level continuity features can be applied before successively higher-level features in order to provide more accurate and more specific scenes. For example, the visual and position continuities can be

used to identify sequential shots showing a high degree of visual similarity in some scenes or in part of the scenes. Most of the current scene-detection algorithms that are based on visual similarity implicitly model the effects of visual and position continuity. However, the visual-similarity method almost always over-segments the video where there are lots of focal movements or other forms of continuity at work. For example, by examining focal distance continuity, we can identify new scenes based on concentration or enlargement rules.

In Sect. 5, we will illustrate the use of the first three continuity features to extract scenes.

5 The detection of scene boundaries

This section describes the details of employing the first three levels of our framework 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 scenes as well, and shots far away will have little effects on the semantic of the current scene.

Our model-based scene-boundary detection method operates at the shot level and consists of the following steps to uncover scenes based on the first three levels of the continuity framework:

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 and Chua (2000). Here we use a combination of black frames, static frames and audio silence to identify the boundaries of commercial blocks. The method has been found to be reliable in filtering out most commercial blocks.
c) We merge the shots into scenes using the visual similarity criteria. This is equivalent to enforcing the visual and position continuity of the framework. We called the resulting set of scenes the scene segments. This approach is able to detect most of the scenes corresponding to simple events that take place in the same location with similar background (Hanjalic et al. 1999), such as those composed using serial content, parallel rule and 180° rule.
d) The above approach tends to over-segment those scenes composed using the concentration or enlargement rules. This is because the sequence of shots appearing in a concentration or enlargement scene is quite different visually. Our next step is therefore to apply the camera's focal distance continuity to identify scenes defined using the more complex cinematic rules, such as those composed using the enlargement or concentration rules.

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

5.1 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 and Chua (2001) to compute shot similarities, and Wang et al. (2001) cluster the group of similar shots into scenes using the tiling window method. The tiling window helps to enforce the locality constraint.

5.1.1 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

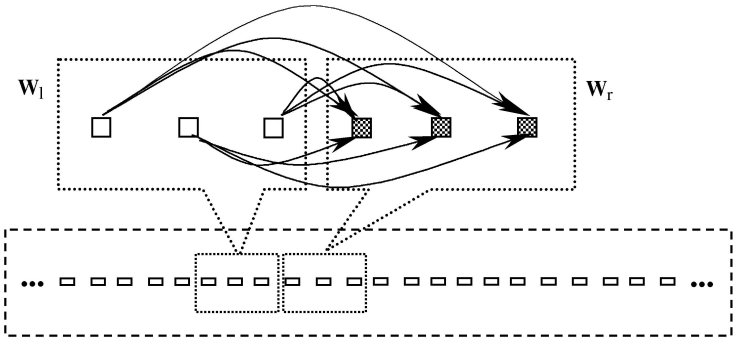


Fig. 7. Similarity comparison between two tiling windows

a sequence of video frames (Shan and 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. (a) We model the content of each frame using three visual feature values, the 1st and 2nd color moments, and the average edge measure modeled based on the number of DCT blocks with high energy values (Chua et al. 2002). These features are all extracted directly from DCT coefficients in MPEG video. (b) We 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 and S_d, is simply the proportion of frame matches between the two shots as follows:

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

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

$$Sim(S_j, S_k) = \sum_{i \in [\text{feature}]} Sim_i(S_j, S_k) \bullet w_i$$

$$\text{where } w_i > 0 \text{ and } \sum_{i \in [\text{feature}]} w_i = 1. \quad (2)$$

The algorithm has been found to be effective in locating similar shots of varying lengths (Chen and Chua 2001).

5.1.2 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 content of the left window are most dissimilar to that on the right (see Fig. 7). 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 in the left and right windows respectively. We first use Eq. (2) to compute the similarities between each pair of shots for each window; that is, we compute all $Sim(S_i, S_j)$ values, where $S_i \in W_l$, $S_j \in W_r$. Next we compute the similarity between W_l and W_r:

$$Sim(W_l, W_r) = \frac{1}{|W_l| \times |W_r|} \sum_{i}^{|W_l|} \sum_{j}^{|W_r|} Sim(S_i, S_j). \quad (3)$$

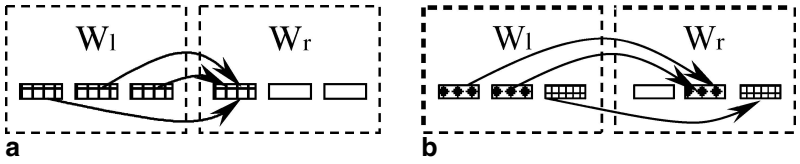


Fig. 8. Tiling windows over a serial and parallel scene

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 and Plaunt 1993), is as follows.

a. 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 $Sim(W_l, W_r)$ using Eq. (3).
b. Plot the sequence of similarity values for 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 the left and right sliding windows for each position. The local minimums on the curve show possible positions of scene boundaries.
c. Use the threshold τ_s to remove those local minimums where the difference between left and right sliding windows is not significant enough to conclude a scene boundary.
d. Select the remaining local minimums, 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 shares many common backgrounds. Hence the majority of the shots belonging to the same scene possess common visual features, which differ from those of other scenes. This is true in most of the simple scenes composed using the serial content rule.

5.1.3 Detection of serial scenes and parallel scenes

The algorithm outlined in Sect. 5.1.1 is designed to handle simple visually similar scenes in which visual and temporal continuity are well preserved and the 180° rule and the parallel rule are obeyed. The shots containing the tiling windows possess a high similarity, as shown in Fig. 8a. They may also contain parallel scenes containing two or more sequences of interleaving shots as shown in Fig. 8b.

5.2 Detection of scenes defined using enlargement, concentration and 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 equivalent to enforcing camera focal distance continuity. 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 the main objects. The focal distance ranges from 6 (extreme long distance shot), through 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 the 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.

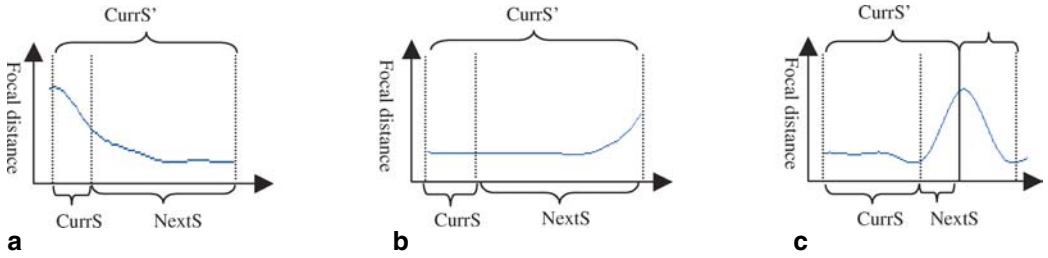


Fig. 9. Merging of scene segments based on cinematic rules

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, then do the following (See Fig. 9 for illustration)
- Case 1 (Concentration rule): If the focal distance of shots reduces 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 distance 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 distance of shots in *CurrS* increase into *NextS*, but the focal distance of the shots in *NextS* exhibit a peak, showing an increasing followed by a decreasing trend, then it indicates that a scene boundary occurs within *NextS*. We divide the *NextS* into two parts separated at the peak (see Fig. 9c). 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*.

(b) Proceed to the next scene segment by setting $CurrS = NextS$, and assigning *NextS* to the following scene segment.

(c) Repeat from Step (a) until all the scene segments have been considered.

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

6 Experiment results and evaluation

We used one full-length movie and two documentaries (one was short and the other was full length) to test our proposed scene analysis method. The movie was obtained from the Media Corporation of Singapore (MediaCorp). It was an hour-long detective TV series including commercials. The movie was more artistically composed and contained many changes in locations where the stories take place. From these videos, we observed clear cinematic rules used to compose the scenes. In order to remove the noise introduced by the commercials, which use different styles of composing the contents, we filtered out the commercials before applying our scene detection algorithm.

In order to test our system objectively, we need to extract the ground truth on the boundaries of the scenes. As the scene is a high-level concept, there will be a certain degree of subjectivity in determining the scene boundary. To avoid such problems, we used two human viewers to view the movie independently and propose scene boundaries. The scenes segmented by the viewers were mostly the same. The differences were resolved through discussions. We then used the final set of boundaries as the ground truth. Table 1 summarizes the statistics of the three test videos. There are altogether 94 scenes in over 70 minutes of video.

Table 2 shows the detailed results of scene boundary detection at the end of first stage, and Table 3 shows the results at the end of second stage. The two stages are (A) after the clustering of visually similar scenes described in Sect. 5.1 and (B) after the application of camera focal distance continuity to identify more complex scenes as described in Sect. 5.2. From the Table 4, we can see that at the end of Stage A, we

Table 1. Statistics of test videos

	Frame #	Shot #	Scene #	Duration
Video 1 (movie)	62 209	521	42	41.5 min
Video 2 (documentary)	4322	26	4	2.9 min
Video 3 (documentary)	39 002	244	48	26 min
Overall	105 533	791	94	70.4 min

Table 2. Results after the Stage A

	Total #	Wrong #	Miss#
Video 1: Movie	61	21	2
Video 2: Documentary	6	2	0
Video 3: Documentary	56	17	9
Overall:	123	40	11

Table 3. Results after the Stage B

	Total #	Wrong #	Miss#
Video 1: Movie	46	6	2
Video 2: Documentary	4	0	0
Video 3: Documentary	48	11	11
Overall:	98	17	13

Table 4. Scenes detected using scene segments vs. after applying cinematic rules

	Total	Wrong	Miss	Precision	Recall
Stage A:	123	40	11	67.5%	88.3%
Stage B:	98	17	13	82.7%	86.2%

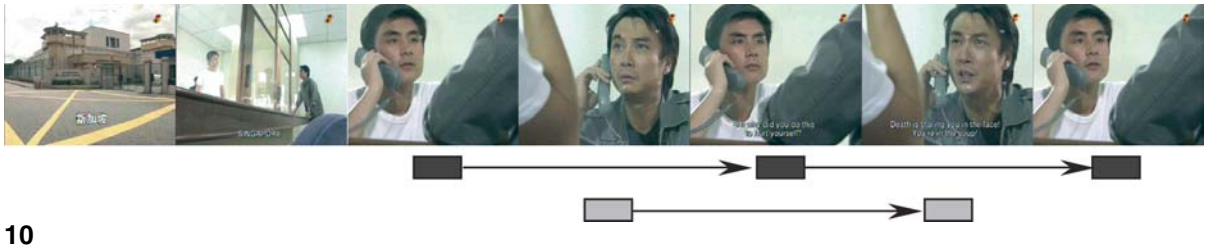
could achieve a high recall of 88.3%, but the precision is quite low at 67.5%. This is to be expected as the tiling window method based on visual similarity criteria tends to over-segment scenes. After the application of the cinematic rules in Stage B, we were able to improve the precision drastically to 82.7%, while the recall drops only slightly to 86.2%. The results clearly demonstrate that the use of cinematic rules based on the concept of continuity is effective. In fact, our preliminary observation suggests that by applying the motion continuity, we can further improve the precision.

To illustrate the results, we present and analyze two more complex scenes correctly extracted by our method. The first example (see Fig. 10) shows a parallel scene involving the conversation between two persons in a room of the jail. There are two links in this parallel scene for main objects, and two shots for the context. The two links are interleaved, clearly showing the presence of simultaneous actions taking place.

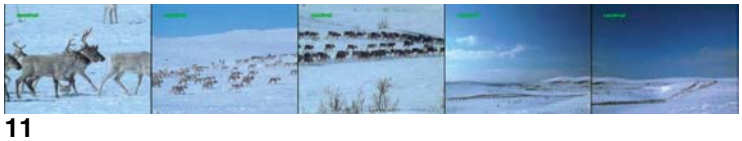
The second example (see Fig. 11) shows a scene of an American caribou migration and the end of this scene. Here the first shot tells us clearly that the protagonist in this scene is caribou. The focal length zooms out gradually into extremely long shot of the environment. Through this process, audience is reminded that caribou live in a cold and snowy environment.

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. In order to unify the cinematic rules, this paper proposes a framework based on the concept of continuity. Starting from a set of shots, the framework successively applies the concept of visual, position, camera focal distance, motion, audio and semantic continuity, to group the shots that exhibits some form of continuity into scenes. We tested the framework by enforcing the first three levels of continuity. This is equivalent to applying the serial, parallel, 180°, and concentration/enlargement/general rules implicitly. We tested our system on three videos of



10



11

Fig. 10. An example of a parallel scene
Fig. 11. An example of an enlargement rule scene

about 70 min in duration. The system was found to be effective.

The work reported here represents only the beginning to this line of research. The framework helps to explain the principles and the heuristics behind most cinematic rules. Further research will be carried out in the following directions. First, as we only implemented the first three levels of continuity in the work reported here, we will consider the use of more complex cinematic rules to characterize complex scenes. In particular, we will investigate the use of motion continuity and audio continuity to improve the precision of scene segmentation. Second, we will investigate formal models for film grammar and other scene semantics based on features like text, audio, shot categories and other domain knowledge, and develop stochastic techniques such as the hidden Markov model (Rabiner 1989) to discover scenes in a learning-based approach. Third, we noticed that scene is a rather fuzzy and subjective concept and different users have different ideas of what the scenes are, thus we are investigating adaptive technique to perform user-oriented scene detection. We will also investigate how user-oriented scene detection can be used to achieve user-oriented video summarization for personal video adaptation.

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. Aoki H, Shimotsuji S, Hori O (1996) A shot classification method of selecting effective key-frames for video browsing. In: Proceedings of the fourth ACM international conference on multimedia, Boston, Mass., 18–22 November 1996. ACM Press, New York
2. Arijon D (1991) Grammar of the film language. Silman-James Press, Los Angeles
3. Arman F, Depommier R, Hsu A, Chiu M-Y (1994) Content-based browsing of video sequences. In: Proceedings of the second ACM international conference on multimedia, San Francisco, Calif., 15–20 October 1994. ACM Press, New York
4. Beaver F (1994) Dictionary of film terms. Twayne Publishing, New York
5. Chen LP, Chua TS (2001) A match and tiling approach to content-based video retrieval. In: Proceeding of the 2001 IEEE international conference on multimedia and expo, Tokyo, Japan, 22–25 August 2001
6. Chua TS, Kankanhalli M, Lin Y (2000) A general framework for video segmentation based on temporal multi-resolution analysis. In: Proceedings of the 3rd international workshop on advanced image technology, Fujisawa, Japan.
7. Chua TS, Ruan LQ (1995) A video retrieval and sequencing system. ACM Trans Inf Syst 13(4):373–407
8. Chua TS, Zhao Y, Mohan K (2002) Detection of human faces in a compressed domain for video stratification. Vis Comput 18:121–133
9. Davenport G, Smith TA, Princever N (1991) Cinematic primitives for multimedia. IEEE Comput Graph Appl 11(4):67–74
10. Eisenstein S (1968) The film sense. Faber and Faber, London

11. Hanjalic A, Lagendijk RL, Biemond J (1999) Automated high-level movie segmentation for advanced video retrieval system. IEEE Trans Circuits Syst Video Technol 9(4):580–588
12. Hari S, Chang S-F (2000) Determining computable scenes in films and their structures using audio-visual memory models. In: Proceedings of ACM Multimedia 2000, Los Angeles, Calif., 30 October–3 November 2000. ACM Press, New York
13. Hearst MA, Plaunt C (1993) Subtopic structuring for full-length document access. In: Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval, Pittsburgh, Pa. ACM Press, New York
14. Jain AK, Vailaya A, Wei X (1999) Query by video clip. Multimedia Syst 7:369–384
15. Kender JR, Yeo BL (1998) Video scene segmentation via continuous video coherence. In: IEEE Computer Society conference on computer vision and pattern recognition. IEEE Computer Society Press, Silver Springs, Md.
16. Koh CK, Chua TS (2000) Detection and segmentation of commercials in video. UROP Report, School of Computing, National University of Singapore
17. Mercer J (1971) An introduction to cinematography. Stipes, Champaign, Ill.
18. Noma T, Okada N (1992) Automating virtual camera control for computing animation. In: Thalmann NM, Thalmann D (eds) Creating and animating the virtual world. Springer-Verlag, Berlin Heidelberg New York
19. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
20. Thompson R (1998) Grammar of the shot. Focal Press, London
21. Rui Y, Huang TS, Mehrotra S (1998) Exploring video structure beyond the shots. In: IEEE international conference on multimedia computing and systems, Austin, Tex., 28 June - 1 July 1998. IEEE Computer Society, Los Alamitos, Calif.
22. Shan MK, Lee SY (1998) Content-based video retrieval based on similarity of frame sequence. In: International workshop on multi-media database management systems, Dayton, Ohio, 5–7 August 1998. IEEE Computer Society Press, Los Alamitos
23. Wang J, Chua TS, Chen L (2001) Cinematic-based model for scene boundary detection. In: Proceedings of the international conference on multi-media modeling, Amsterdam, 5–7 November 2001
24. Yeung M, Liu B (1995) Efficient matching and clustering of video shots. In: Second international conference on image processing, Washington, D.C., 23–26 October 1995. IEEE Computer Society Press, Los Alamitos, Calif.
25. Yeung M, Yeo BL, Liu B (1996) Extracting story units from long programs for video browsing and navigation. In: Proceedings of the international conference on multimedia computing and systems, Hiroshima, 17–23 June 1996. IEEE Computer Society Press, Los Alamitos, Calif.
26. Yoshitaka A, Ishii T, M. Hirakawa M, Ichikawa T (1997) Content-based retrieval of video data by the grammar of film. In: Proceedings of the IEEE Symposium on Visual Languages. IEEE Computer Society Press, Los Alamitos, Calif.
27. Zhong D, Zhang H, Chang SF (1996) Clustering methods for video browsing and annotation. In: Storage and retrieval for still image and video database, vol 4. SPIE, Bellingham, Wash.

JIHUA WANG received his BSc degree in computer science from the Beijing Institute of Technology (1996), and MSc in computer science from the Institute of Computing Technology, Chinese Academy of Sciences (1999). He is presently a research scholar and PhD candidate in the School of Computing, National University of Singapore. His main research interests are in image retrieval, video analysis, video retrieval and related applications.

TAT-SENG CHUA was the acting and founding dean of the School of Computing, National University of Singapore, from 1998 to 2000. He spent three years as a research staff member at the Institute of System Science (now LIT) in the late 1980s. His main research interest is in multimedia information processing and, in particular, video and text retrieval and information extraction. He leads a large-scale research project to retrieve digital video, and web-based information. Dr. Chua has organized and served as program committee member of numerous international conferences in the area of computer graphics and multimedia. He serves on the editorial boards of IEEE Transactions of Multimedia (IEEE), The Visual Computer (Springer-Verlag), and Multimedia Tools and Applications (Kluwer). He obtained his PhD from the University of Leeds, UK.