A Framework for Video Scene Boundary Detection

Jihua Wang
School of Computing, NUS
Singapore, 117543
65 - 68744362
wangjihu@comp.nus.edu.sg

Tat-Seng Chua
Associate Professor
School of Computing, NUS
Singapore, 117543
65 - 68744362
chuats@comp.nus.edu.sg

ABSTRACT

Most current video retrieval systems use shot as the basis for information organization and access. In cinematography, scene is the basic story unit that the directors use to 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, cameral focal distance, motion, audio and semantic continuity to group shots that exhibit some form of continuity into scenes. The idea is tested using the first three levels of continuity to extract the scenes defined using most common cinematic rules. The method has been found to be effective.

Categories and Subject Descriptors

H.30 [**Information Storage and Retrieval**]: General; H.5.1 [**Information Interface and Presentation**]: Multimedia Information System. 1.2.4.

General Terms

Algorithms, Experimentation, Theory.

Keywords

Cinematic model, scene detection, video retrieval.

1. INTRODUCTION

Most current video retrieval systems use shots as the basis to organize video contents (Yeung & Liu 1995, Zhong et al 1996). But viewers see and remember video in terms of events, episodes and stories. Here, we use the term "scene" prevailing in cinematography to denote episode and story.

Scene consists of a small number of interrelated shots that are unified by location or dramatic incident (Beaver 1994). Some cinematic rules including the 180° rule (Thompson 1998, Hari & Chang 2000), montage rules (Eisenstein 1968) and other cinematic continuity rules were developed to convey director's idea in a consistent way.

The main objective of this research is to use the cinematic rules as the basis to emulate the creative process of human directors in composing video. 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. Section 4 presents our overall framework based on the concept of the content continuity. Section 5 presents the computational procedure to perform the scene boundary detection. The results of experiment are discussed in Section 6. Finally, Section 7 concludes the paper.

2. RELATED WORKS

In order to derive higher-level semantic entities, a number of recent works investigated the extraction of 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 et al 1996, Zhong et al 1996, Rui et al 1998). 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 minimums in this continuous measure to detect scene boundaries. These techniques are able to handle only simple scenes containing shots that share high visual similarity.

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) on the other hand used the idea of linking similar shots together into treads, and created scenes that consisted of shots coming mostly from one or more interleaving treads. 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. 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

Multimedia'02, December 1-6, 2002, Juan-les-Pins, France.

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/ general rules.

3.1 CINEMATIC MODEL for VIDEO SCENE COMPOSITION

This section provides an overview of the 180° rule and the set of montage rules (Davenport et al 1991, Chua & Ruan 1995).

- The 180° rule, also called Triangle Principle (Arijon 1976, Thomposon 1998, Hari and Chang 2000), states that the camera in the shooting should stay in one side of the line formed by the main subjects. It ensures that the relative positions of the subjects on the screen within a scene are unchanged (see Figure 1).

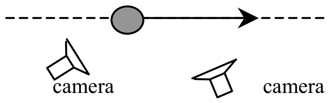

Figure 1. 180° rule

- Parallel rule: The rule is frequently used to model the subjects' interactions. The shots with different themes are shown alternately (Figure 2). It is frequently used to model interactions between two parties such as conversation, hunting, chasing etc.

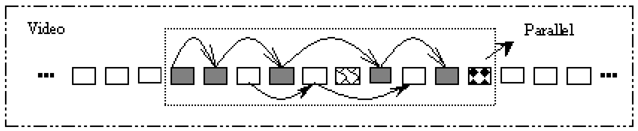

Figure 2. Parallel rule, interaction of two subjects

- Concentration rule: It starts with the long distance shot, and zooms into close up shots of the main objects to introduce the main objects and their context.
- Enlargement rule: It is the reverse of the concentration rule. It is also used to show the main objects and the environment.
- General rule: It is the combination of concentration followed by enlargement rules
- 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.

4. THE OVERALL FRAMEWORK for SCENE DETECTION

The theories of montage and cinematic rules have been used by the directors and editors to create coherent stories. From our analysis, it can be seen that continuity is an invariant theme that unifies all cinematic rules. For example, visual similarity 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/enlargement rules. In addition, view, motion or even audio continuities are also used based on some general or specific rules.

To capture the concept of continuity, we develop a framework for scene boundary detection. Figure 3 shows the continuity framework. The relations between each level of continuity and cinematic rules are explained below.

- Visual continuity: It exists between successive shots with similar background.
- Position continuity: It ensures that objects preserve their relative positions in the scene.
- Focal distance continuity: It typically changes in a continuous way in the establishing stage as well as at the ending stage of a scene.
- Object motion continuity: The direction of motion of the main object within a scene should be the same. Figure 4 shows a two-person chasing scene in the four consecutive shots in which the two men all run from left to the right.
- Audio continuity: The sound track in a movie contains environment sound and dialogue. The same scene should possess similar environment sound and dialogue by the same speakers. This rule might not be true for other types of scenes 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 composed based on semantic coherence, such as news video etc. It is a high-level concept that is content and domain dependent.

The framework is designed in such a way that the lower layer continuity features can be applied before successively higher layer features to provide more accurate and more specific scenes. Most of the current scene detection algorithms that are based on visual-similarity implicitly model the effects of visual and position continuity to detect the scenes. However, the visual similarity based method almost always over-segments the video where there are lots of focal movements or other higher-level forms of continuity at work. For example, by examining focal distance continuity, we can observe new scenes based on concentration/enlargement/general rules.

By using the framework, we want to use the concept of continuity at different level to detect scenes in a video to different degree of sophistication. In the following section, we will illustrate the use of the first three continuity features to extract scenes.

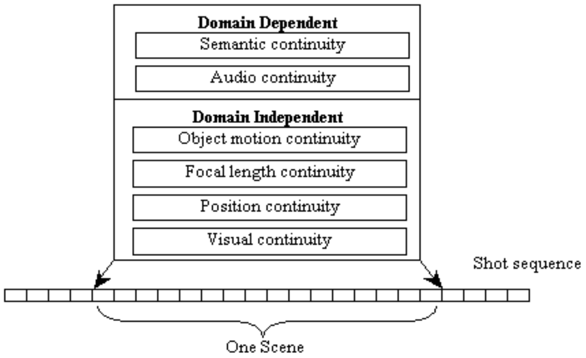

Figure 3. Content continuity

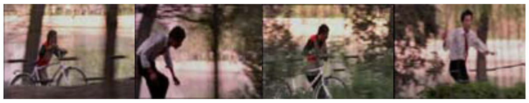

Figure 4. Motion continuity is preserved in the four consecutive shots. (Beijing Bicycle, China)

5. SCENE BOUNDARY DETECTION

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 3 level of continuity framework: (a) We segment the video into shots by employing the TMRA method (Chua et al 2000). (b) We filter out the commercials using the method developed in Koh & Chua (2000); (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. (d) Finally, we apply the camera focal distance continuity to identify scenes defined using more complex cinematic rules composed using the enlargement/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

We consider shot similarity comparison as the comparison between two sequences of frames. We want to support efficient matching of both exact and partially similar shots, and also take into consideration the temporal variations across the entire shot. We employed the technique developed in Chen & Chua (2001) to: (a) We model the content of each frame using three visual feature values: the 1st and 2nd color moments, and the average edge measure modelled based on the number of DCT blocks with high energy value (Chua, Zhao & Mohan 2002). (b) We model the content of the entire shot as the trajectories of these three quantized feature values. Using these features, we use the algorithms presented in Chen & Chua (2001) to determine the shot visual similarity. We then use the overall sliding window algorithm (Wang et al 2001), which is similar to the text tiling method (Hearst & Plaunt 1993), to detect the possible scene boundaries.

The algorithm is effective in handling simple visually similar scenes in which visual and position continuity are well preserved. It also naturally models the 180° rule as well as parallel rule because the shots within the tiling windows posses high similarity as shown in Figure 5a. It also captures parallel scenes that contain two or more sequences of inter-leaving shots as shown Figure 5b.

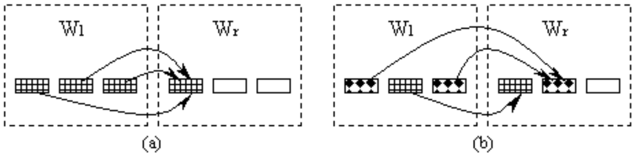

Figure 5. Tiling window over serial and parallel scene

5.2 Detection of Scenes 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/enlargement/general 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 main objects. The focal distance ranges from 6 (extreme long distance shot) to 1 (close-up shot or equivalent to showing the face of a person on half the screen).

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 τ_s, then: (See Figure 6 for illustration)

- Case 1 (Concentration rule, Figure 6a): If the focal distance of the shots reduces steadily from *CurrS* into *NextS*, the *CurrS* is merged with *NextS*.
- Case 2 (Enlargement rule, Figure 6b): If the focal distance of the shots increases from *CurrS* into *NextS*, then merge the *CurrS* and *NextS*.
- Case 3 (General rule): If the focal distance of the shots in *CurrS* increases into *NextS* and exhibits a peak in *NextS*, followed by decreasing trends. We divide the *NextS* into two parts separated at the peak (see Figure 6c). 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.

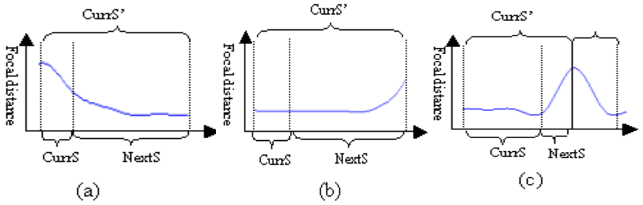

Figure 6. Merging of scene segments based on cinematic rules

6. RESULTS and EVALUATION

We use one full-length movie and two documentaries to test our proposed scene analysis method. From these videos, we can observe clear cinematic rules used to compose the scenes. In order to remove the noise introduced by the commercials, we filtered out the commercials before applying our scene detection algorithm.

In order to test our system objectively, we used two human viewers to view the videos independently and propose scene boundaries. The scenes segmented by the viewers are mostly the same and the differences are resolved through discussions.

Table 1 summarizes the statistics of the three test videos. There are altogether 94 scenes in over 70 minutes of video.

Table 1. Statistics of test videos

	Frame #	Shot #	Scene #	Duration
Movie	62,209	521	42	41.5 min
Documentary 1	4322	26	4	2.9 min
Documentary 2	39,002	244	48	26 min
Overall:	105,533	791	94	70.4 min

From Table 2, we can see that at the end of Stage A, we 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, especially those complex scenes composed using the concentration, enlargement and general rules. We see, however, that after the application of cinematic rules in Stage B, we are 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 is effective.

Table 2. **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%

Stage A: results after applying techniques in Section 5.1
Stage B: results after applying techniques in Section 5.2

7. CONCLUSION and FURTHER RESEARCH

In order to unify the cinematic rules, this paper proposes a framework based on the concept of continuity. The framework successively applies the concept of visual, position, cameral focal distance, motion, audio and semantic continuity to group the shots that exhibits some form of continuity into scenes. We test the framework by enforcing the first three level of continuity. This is equivalent to applying the 180° rule, serial rule, parallel rule, and concentration/enlargement/general rule implicitly. We test our system on three videos of about 70 minutes in duration. The system has been 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, 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 the user-oriented scene detection can be used to help us to achieve user-oriented video summarization for personal video adaptation.

8. ACKNOWLEDGMENTS

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.

9. REFERENCES

[1] Arijon, D. Grammar of the Film Language, Silman-James Press LA, 1991.

[2] Beaver, F. Dictionary of Film Terms. Twayne Publishing NY, 1994.

[3] Chen, L.P., Chua, T.S. (2001) A Matched Tiling Approach to Content-based Video Retrieval. Proc. of ICME’2001 (Tokyo, Japan, Aug. 2001).

[4] 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), 119 – 124.

[5] Chua, T.S., Ruan, L.Q. A Video Retrieval and Sequencing System, ACM Trans. Inf. System 13(4) (1995), 373 – 407.

[6] Chua, T.S., Zhao, Y., Kankanhalli, M. Detection of Human Faces in a Compressed Domain for Video Stratification, Visual Computer 18 (2002), 121 – 133.

[7] Davenport, G., Smith, T.A., Princever, N. Cinematic Primitives for Multimedia, IEEE Computer Graphics and Applications 11(4) (1991), 67 – 74.

[8] Eisenstein, Eergei, M. The Film Sense (1968), Faber and Faber Ltd.

[9] 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), 580 – 588.

[10] Hari, S., Chang, S.-F. Determining Computable Scenes in Films and their Structures using Audio-Visual Memory Models. Proc. ACM MM’2000, 95 - 104

[11] Hearst, M.A., Plaunt, C. Subtopic Structuring for Full-Length Document Access. ACM SIGIR (1993), 59 – 68.

[12] Kender, J.R., Yeo, B.L. Video Scene Segmentation via Continuous Video Coherence. Proc. IEEE Int’l Conf. On CVPR (1998), 367 – 373.

[13] Koh, C.K., Chua, T.S. Detection and Segmentation of Commercials in Video, UROP Report (2000), School of Computing, National University of Singapore.

[14] Rabiner, L. R. A Tutorial On Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE 77(2) (1989), 257 – 286.

[15] Rui, Y., Huang, T.S., Mehrotra, S. Exploring Video Structure Beyond the Shots. Proc. IEEE Conf. on Multimedia Computing and Systems (1998), 237 – 240.

[16] Thompson, R. Grammar of the Shot (1998). Focal Press.

[17] Yeung, M., Liu, B. Efficient Matching and Clustering of Video Shots, Proc. IEEE ICIP’ 95 1 (1995), 338 – 341.

[18] Yeung, M., Yeo, B.L., Liu, B. Extracting Story Units from Long Programs for Video Browsing and Navigation, IEEE Proc. of Multimedia’96 (1996), 296 – 305.

[19] L. R. Rabiner (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE 77(2) 257 – 286.

[20] Wang, J., Chua, T.S., Chen, L.P. Cinematic-based Model for Scene Boundary Detection. Proc. MMM’01 (Multimedia Modelling 2001), 3 – 18.

[21] 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), 310 – 317.

[22] 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), 239 – 246.