INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11/ m3110
MPEG 97
February 1998/San José

Title: **Video Analysis and Representation**

Source: **Thomas S. Huang, Yong Rui, Trausti Kristjansson, Milind Naphande and Yueting Zhuang**
University of Illinois at Urbana-Champaign

Status: **Proposal**

1. Introduction

Recent years have seen a rapid increase of the usage of multimedia information. Of all the media types (text, image, graphic, audio and video), video is the most challenging one, as it combines all the other media information into a single data stream. Owing to the decreasing cost of storage devices, higher transmission rates, and improved compression techniques, digital video is becoming available at an ever increasing rate.

Because of its length and unstructured format, efficient access to video is not an easy task. From the perspective of browsing and retrieval, video is analogous to a book. Access to a book is greatly facilitated by a well designed table of content (TOC) which captures the semantic structure of the book. For current existing video, a lack of such a TOC makes the task of browsing and retrieval very inefficient, where a user searching for a particular object of interest has to use the time-consuming ``fast forward" and ``rewind" operations. Efficient techniques need to be developed to construct video TOC to facilitate user's access.

Before we explore such techniques, it is worth while to formalize the terminologies used in video analysis. Ideally, video has a well defined hierarchy consisting of *video*, *scene*, *shot*, and *key frame* levels. This video hierarchy is illustrated in Figure 1.

Raw *video* is an unstructured data stream, consisting of a sequence of video shots. A *shot* is an unbroken sequence of frames recorded from a single camera, which forms the building block of a video. It is a physical entity and is delimited by shot boundaries. Since shot boundaries exist physically, automatic shot boundary detection is possible. In many video applications, it is too time-consuming for the user to view the entire set of shots of the whole video, and *key frames* are used to facilitate quick browsing. *Key frame* is the frame which can represent the salient content of the shot. Depending on the
content complexity of the shot, one or more key frames can be extracted from a single shot.

To model the video structure at a semantic level, an abstraction of scene is introduced. *Scene* is defined as a collection of semantically related and temporally adjacent shots, depicting and conveying a high-level concept or story. While *shot* is the building block of video, it is *scene* that conveys the semantic meaning of the video to the viewers. When we watch a video, we never concentrate on how

shots are changed but rather we concentrate on how the story is developed. The discontinuity of shots is overwhelmed by the continuity of a scene [1].

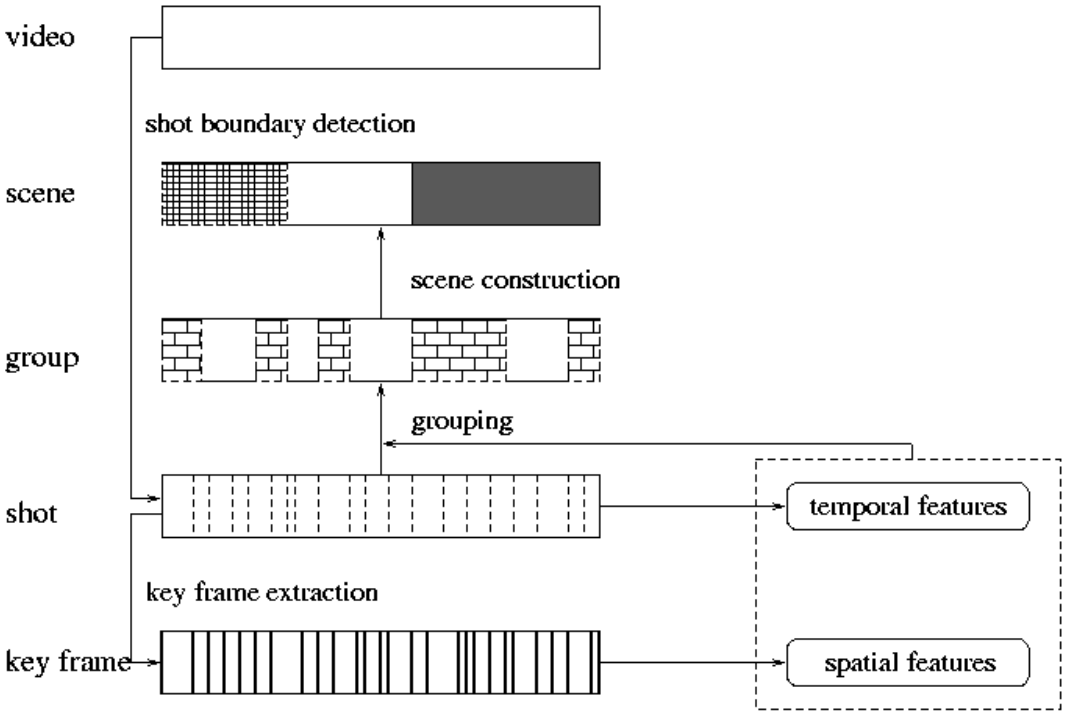


Figure 1. Video structure

While shots are marked by physical boundaries, scenes are marked by semantic boundaries (*Some of the early literatures in video parsing misused the phrase scene change detection for shot boundary detection. But as we can see, these two concepts are very different. To avoid any later confusion, we will use shot boundary detection for the detection of physical shot boundaries while using scene boundary detection for the detection of semantic scene boundaries.*).

In summary, the video hierarchy contains five levels (video, scene, group, shot, and key frame, where group is an intermediate entity), from top to bottom decreasing in unit length. The purpose of constructing video TOC is to convert an unstructured raw video into the above structured video hierarchy to assist user's access.

Over the past few years, progress has been made in shot boundary detection and key frame extraction, which are the bases for later scene construction. Although important, shot and key frame are not closely related to the semantics of the video and normally has large number of entries. They are still difficult for the user to use. It is not uncommon that a modern movie contains a few thousand shots and key frames. This is evidenced in [2] -- there are 300 shots in a 15-minute video segment of the movie ``Terminator 2 - the Judgment Day'' and the movie lasts 139 minutes. Because of the large number of key frames, a simple 1D array presentation of key frames for the underlying video is almost meaningless. More importantly, people watch the video by its semantic scenes not the physical shots or key frames. Shots can not convey meaningful semantics unless they are purposely organized into scenes. The video TOC construction at the scene level is thus of fundamental importance to video browsing and retrieval.

In section 2, we will briefly review and evaluate existing techniques in shot boundary detection and key frame extraction, as well as presenting our approaches. The construction of TOC at the scene level is

discussed in section 3. Section 4 presents our future work in more accurate scene structure construction based on multiple media.

2. Shot Boundary Detection and Key Frame Extraction

2.1 Shot boundary detection

In general, automatic shot boundary detection techniques can be classified into five categories, i.e. pixel based, statistics based, transform based, feature based, and histogram based. *Pixel based* approaches use the pixel-wise intensity difference as the indicator for shot boundaries [3,4]. One of its drawbacks is its sensitivity to noise. To overcome this problem, Kasturi and Jain propose to use intensity statistics (mean and standard deviation) as the shot boundary detection measure. Exploring how to achieve faster speed, Arman, Hsu and Chiu propose to use the DCT coefficients in the compressed domain as the boundary measure. Other transformed based shot boundary detection approaches make use of the motion vectors, which are already embedded in the MPEG stream. Zabih et al. address the problem from another angle. The edge features are first extracted from each frame. Shot boundaries are then detected by comparing the edge difference. So far, histogram difference is the most popular approach used in shot boundary detection. Several researchers claim that it achieves good trade-off between accuracy and speed [3]. Two comprehensive comparisons of various shot boundary detection techniques are in [5,6].

One of the problems of the above approaches is that they use a predefined threshold to determine the similarity or dissimilarity of successive frames. The determination of the threshold is not always easy. Based on the approach developed by Gunsel, Ferman and tekalp[13], we proposed, together with Kodak, an unsupervised clustering based approach which effectively avoid this [12]. The flow chart of the approach is the following:

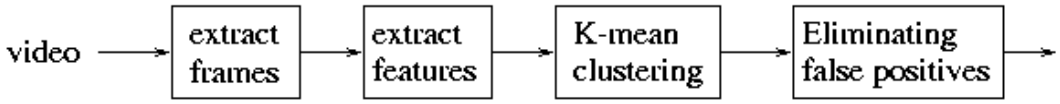


Figure 2. The processing flow chart

In this approach we use both the histogram difference and pixel difference as the features to do the shot segmentation. They are complementary features and thus result in good performance. Some experimental results of this approach over real-world videos are illustrated below:

Sequence	# of frames	# of shots	Shots detected	False alarms
1	3975	66	66	3
2	2415	32	32	0
3	2136	39	37	1
4	2715	53	52	7

The comparison of this approach and existing approaches is shown in Table 2.

Algorithm	% accuracy	% false alarms
Our approach	98.5	5.79
Histogram difference	88	10.53
Compressed domain	59	71.05

2.2 Key frame extraction

After the shot boundaries are detected, corresponding key frames can then be extracted. Simple approaches may just extract the first and last frames of each shot as the key frames. More sophisticated key frame extraction techniques are based on shot activity indicator and shot motion indicator.

2.2.1 Shot boundary based approach

After the video streams is segmented into shots, a natural and easy way of key frame extraction is to use the first frame of each shot as the shot's key frame. Although simple, the number of key frames for each shot is limited to one, regardless of the shot's visual complexity. Furthermore, the first frame normally is not stable and does not capture the major visual content.

2.2.2 Visual content based approach

Zhang et. al. propose to use multiple visual criteria to extract key frames.

Shot based criteria: The first frame will always be selected as the first key frame; but, whether more than one key frame need to be chosen depends on other criteria.

Color feature based criteria: The current frame of the shot will be compared against the last key frame. If significant content change occurs, the current frame will be selected as a new key frame.

Motion based criteria: For a zooming-like shot, at least two frames will be selected: the first and last frame, since one will represent a global, while the other will represent a more focused view. For a panning-like shot, frames have less than 30\% overlap are selected as key frames.

2.2.3 Motion analysis based approach

Wolf proposes a motion based approach to key frame extraction [7]. He first computes the optical flow for each frame, and then computes a simple motion metric based on the optical flow. Finally he analyzes the metric as a function of time to select key frames at the local minima of motion. The justification of this approach is that in many shots, the key frames are identified by *stillness* -- either the camera stops on a new position or the characters hold gestures to emphasize their importance.

2.2.4 Shot activity based approach

Motivated by the same observation as Wolf's, Gresle and Huang propose a shot activity based approach [8]. They first compute the intra- and reference histograms and then compute an activity indicator. Based on the activity curve, the local minima are selected as the key frames.

2.2.5 Summary

Ideally, key frames should capture the semantics of a shot. However, at current stage, the Computer Vision techniques are not advanced enough to automatically generate such key frames. Instead, we have to base key frame selection on low level visual features, such as color, texture, shape of the salient object in a shot. It is obvious that if a frame is important, the camera will focus more on this frame. This is the basic assumption that we use in our clustering based key frame extraction technique [10]. That is, If some visual content is important, there will be more frames having this content. Therefore, if a frame cluster's size is big enough, it deserves a key frame.

Our clustering based key frame extraction approach is not only efficient to compute, it also effectively captures the salient visual content of the video shots. For low-activity shots, it will extract less key frames or one single key frame at most of the time while for high-activity shots, it will automatically extract multiple key frames depending on the visual complexity of the shot. Examples of such cases are illustrated below. Figure 3 (1)(2) shows the two key frames from shot-17 of "Total Recall" because of its visual complexity, while Figure 3 (3)(4) shows the single key frame extraction from "Bridge of Madison County".

Figure 3. Example key frames

3. Scene Construction

To provide the user with better access to the video, construction of a video TOC at a semantic level is needed. Approaches to scene based video TOC construction can be classified into two categories, *model-based* and *general purpose*. In model-based approach, an *a priori* model of a particular application or domain is first constructed. Such a model specifies the scene boundary characteristics, based on which the unstructured video stream can be abstracted into a structured representation. The theoretical framework of this approach was proposed by Swangberg, Shu and Jain in, and it has been successfully realized in many interesting applications, including News Video parsing and TV Soccer program parsing. Since the video parsing is based on the domain model, this approach normally achieves high accuracy. One of the drawbacks of this approach, however, is that for each application a domain model needs to be constructed before the parsing process can proceed. The modeling process is time consuming and requires good domain knowledge and experience.

Another approach to scene based video TOC construction does not require an explicit domain model. Two of the pioneering works of this approach are from Princeton University [1,2] and Toshiba Corp. [9]. In [1,2], the video stream is first segmented into shots. Then time-constrained clustering is used to construct visually similar and temporally adjacent shot clusters. Finally a *Scene Transition Graph* is constructed based on the clusters and *cutting edges* are identified to construct the scene structure. In [9], instead of using Scene Transition Graph, the authors group shots of alternating patterns into scenes (they call *acts*). A 2D presentation of the video structure is then created, with scenes displayed vertically and key frames displayed horizontally.

This video TOC provides the user a much more meaningful way of accessing the video content. The advantages of scene based video TOC over the other approaches are:

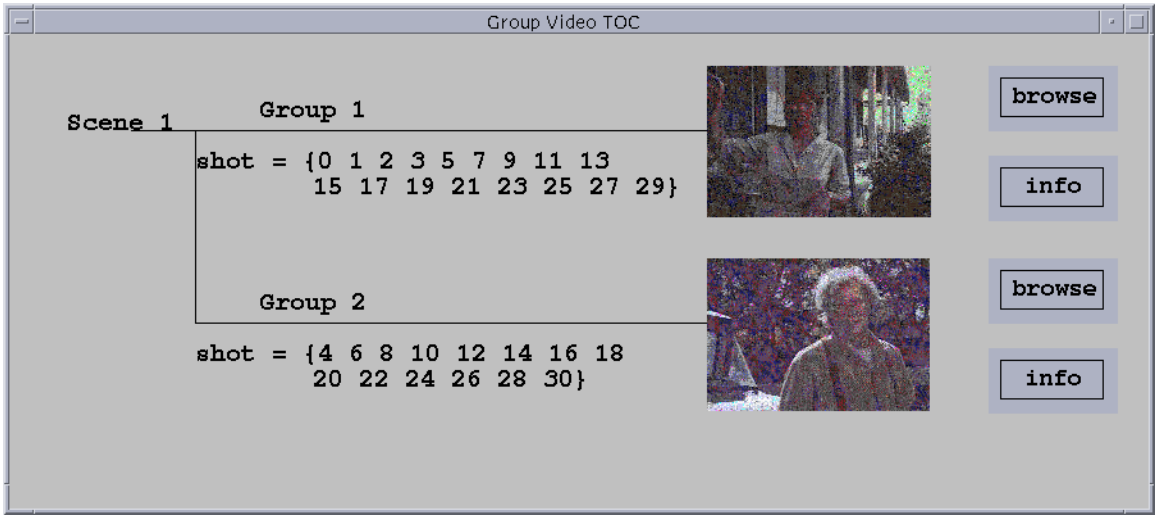
- The other approaches produce too many entries to be efficiently presented to the viewer.
- Shots, key frames, etc convey only physical discontinuity while scenes convey semantic discontinuity, such as scene change in time and/or location.

Our proposed approach [11] to video TOC construction consists of four modules: shot boundary detection and key frame extraction, spatio-temporal feature extraction, time-adaptive grouping, and scene structure construction. Its advantages over existing approaches are: temporal continuity, direct merging to a scene, and on-line processing. The following table summarizes the results of this approach over real-world videos: Bridge of Madison County (BMC)(Romantic-slow), Pretty Woman (PW)(Romantic-fast), Grease (GR)(Music), The Mask (MS)(Comedy), Star Trek (ST)(Science fiction – slow), Star War (SW)(Science fiction – fast), and Total Recall (TR)(Action).

Movie name	frames	shots	groups	Detected scenes	False negatives	False positives
BMC	21717	133	27	5	0	0
PW	27951	186	25	7	0	0
GR	14293	84	13	6	1	0
MS	35817	195	28	12	1	2
ST	18362	77	10	6	0	0

SW	23260	180	31	21	1	10
TR	35154	329	65	21	1	2

An example video TOC is illustrated in the following figure.



4. Extracting semantic representations

4.1 A semantic framework, the OSA representation.

The previous analysis methods have used cinematographic characteristics to find structure in video, as well as some content related visual features of the video for determining salient images, i.e. the key frames. These methods are useful for producing structures that relate to a person's semantic constructs, e.g. TOC. However, these methods do not work with concepts that parallel the semantic constructs a person uses. In the following proposed method the units that are extracted are at a semantic level.

In order to build a semantic representation of a video, the following constructs are identified:

- Objects: e.g. people, Joe, cars, pink, houses etc.
- Activities or events: e.g. talk, walk dance, fight, crash etc.
- Sites: i.e. the spatial framework within which objects are transported through time by events: e.g. inside, kitchen, concert hall, outside, field, city street etc.

These constructs are used to build a machine level representation of a video, called an Object, Site, Activity representation or OSA representation. This representation can be viewed in the above proposed TOC form or can be queried at a semantic level, e.g. "find a car crash".

4.2. Identification of constructs

The proposed method requires recognition of objects in the video. The state of the art in face recognition shows that this is a difficult task from visual features alone. We propose a method that uses all available features, i.e. video, audio and closed caption as well as top down world knowledge. We propose to merge features from all these modalities in a probabilistic framework.

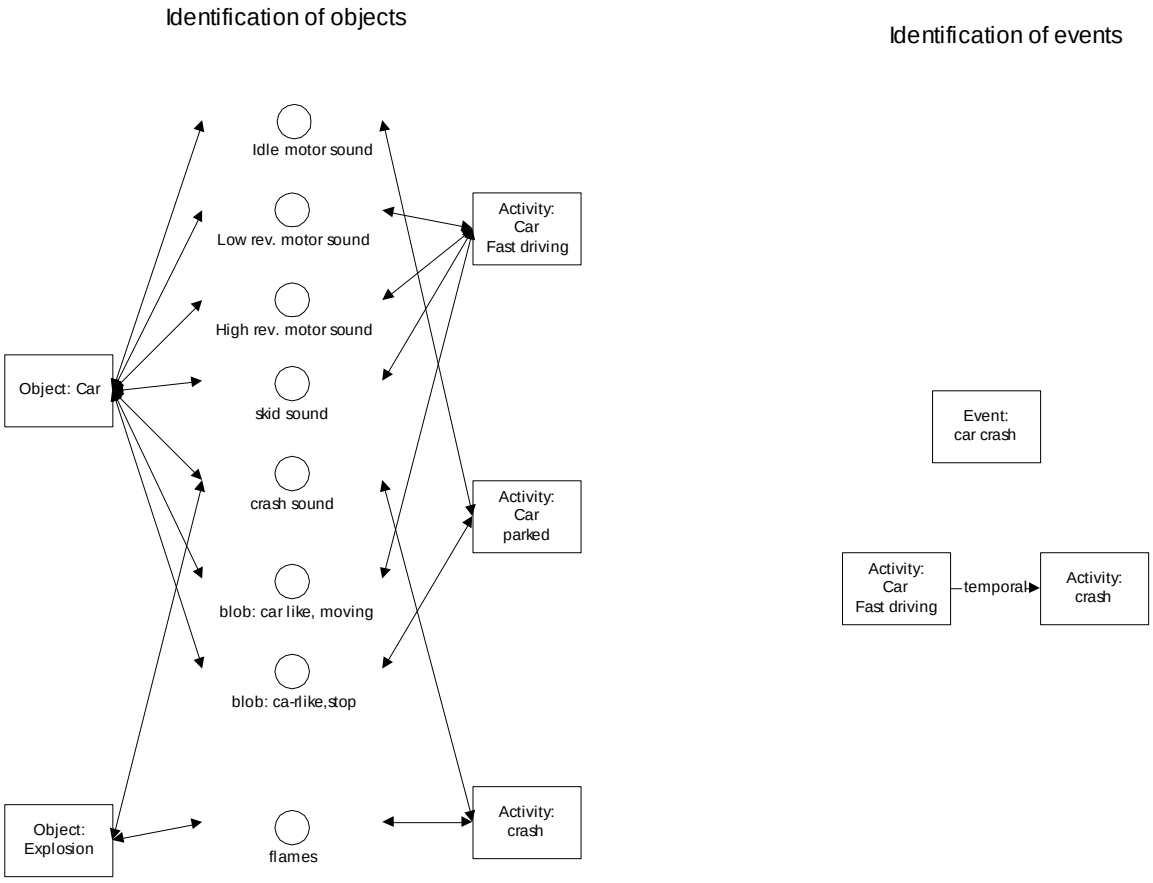
4.2.1 Objects

Objects are recognized by their visual and auditory features. An additional complication is that objects form conceptual classes (e.g. road vehicles). Classes have essential features (e.g. wheels) and incidental features (e.g. four wheels). Specific instances also have characteristic features (e.g. characteristic voice) and incidental features (e.g. blue shirt).

4.2.1 Activity/Events
Events involve objects. Audio is useful for identifying the type of activity or event, e.g. talked, swam, crash, explode, kicked. Motion can be used to support the identification of the event or to relate the activity to a blob in the video.

4.2.3 Sites
Sites are frameworks within which events happen to objects. Sites can be determined by the auditory and visual background. They also constrain which objects are probable. As an example, in a nature scene, the sound of a stream, the rustling of wind in leaves of trees, a blue sky and birds sounds are probable.

4.3 Construction of a hierarchical video representation.
As discussed before a scene can be recognized by grouping shots by their visual features. A scene is at a conceptual level. It is usually spatially and temporally contiguous. Therefore, the objects and site should not change drastically. An integral part of the OSA representation is a scene structure. This structure is determined from the change in site or identified objects. It can therefore be used to construct the scene hierarchy.
The information embodied in the OSA representation can be used to expand the information presented by key frames, e.g. by listing the identified objects or site and generating a written description of the salient event. It can also be used as an additional input to the selection process of key frames.



References:

[1]. Ruud M. Bolle, Boon-Lock Yeo, and Minerva M. Yeung. Video Query: Beyond the keywords. Technical report, IBM Research Report, Oct 17 1996

[2]. Minerva Yeung, Boon-Lock Yeo, and Bede Liu. Extracting story units from long programs for video browsing and navigation. In Proc. IEEE Conf. On Multimedia Computing and Systems, 1996

[3]. HongJiang Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic Partitioning of full-motion video. ACM Multimedia Systems, 1(1), 1993

[4]. A. Hapmpapur, R. Jain, and T. Weymouth. Digital video segmentation. In Proc. ACM Conf. On Multimedia, 1994

[5]. John S. Boreczky and Lawrence A. Rowe. Comparison of video shot boundary detection techniques. In Proc. Of SPIE Conf. On Vis. Commun. And Image Proc. 1996

[6]. Ralph M. For, Craig Robson, Daniel Temple, and Michael Gerlach. Metrics for scene change detection in digital video sequences. In Proc. IEEE Conf. On Multimedia Computing and Systems, 1997

[7]. Wayne Wolf. Key frame selection by motion analysis. In Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc. 1996

[8]. P. O. Gresle and T. S. Huang, Gisting of video documents: A key frame selection algorithm using relative activity measure. In The 2nd Int. Conf. On Visual Information Systems, 1997

[9]. Hisashi Aoki, Shigeyoshi Shimotsuji, and Osamu Hori. A shot classification method of selecting effective key frames for video browsing. In Proc. ACM Conf. On Multimedia 1995

[10].Yueting Zhuang, Yong Rui, Thomas S. Huang and Sharad Mehrotra, Key Frame Extraction by Unsupervised Clustering, submitted to ICIP98

[11].Yong Rui, Thomas S. Huang and Sharad Mehrotra, Constructing Table-of-Content for the Videos, submitted to ACM Multimedia Journal

[12]. M. Naphade, R. Mehrotra, A. M. Ferman , J. Warnick, T. S. Huang, A. M. Tekalp, A High Performance Algorithm for Shot Boundary Detection using multiple cues, submitted to ICIP'98

[13]. B. Gunsel, A. M. Ferman, and M. Tekalp, Video Indexing Through Integration of Syntactic and Semantic Features, Proc WACV'96