STRATIFICATION APPROACH TO MODELING VIDEO

Tat-Seng Chua, Liping Chen and Mohan Kankanhalli
School of Computing
National University of Singapore,Singapore 119260
{chuats, chenlp, mohan}@comp.nus.edu.sg

Last few years have seen an explosion of audiovisual information from many sources around the world. There is an urgent need to develop advanced modeling tool to manage the vast quantity of video information effectively. In this research, we investigate the use of stratification approach to represent the contextual information of video contents as multi-layered strata. Each stratum models the multiple occurrences of a simple concept in the video stream. Strata may overlap and thus the meaning of video at any instance is simply the integration of meanings of all strata present. By using the strata as the base, we can develop advanced functionalities to support flexible retrieval and content-based browsing of video. A prototype has been developed to support the whole process of video management from indexing, retrieval to browsing. The prototype is tested in the domain of News video and the results have been found to be highly satisfactory.

1. Introduction

The wide spread acceptance of MPEG video standards [MPEG 93] has accelerated the use of digital video in our daily life. The use of digital video permits seamless integration of broadcast, telecommunication and computer in a single framework. Such integration facilitates the development of advanced innovative applications that will have profound effects in our everyday life. One such application is the personalized news broadcast that enables the users to view their preferred news summaries, and able to browse quickly to regions of news of their interests.

To support the above application, the video data must be properly modeled and indexed. As video is a complex temporal medium that captures both the sights and sounds of the real world, the modeling of semantic contents of video is a complex task. There are generally two approaches to modeling video data. They are the segmentation approach [Rubin & Davenport 89], and the stratification approach [Aquierre-Smith & Pincever 91]. In the segmentation approach, the video is divided into atomic units called shots. Each shot consists of a visually continuous sequence of frames. The content of each shot is then described individually. To provide the necessary context information, a concept structure is normally superimposed on top of the set of shots. The two-layered structure is used to support effective modeling and retrieval of video sequences based on users' queries [Chua & Ruan 95]; cinematic rules are used to sequence the retrieved video shots into a coherent sequence for presentation.

The stratification approach, on the other hand, focuses on segmenting the contextual information into chunks, rather than dividing physically contiguous frames into shots as is traditionally done. Each chunk is known as a *stratum*, which describes the occurrences of a simple concept like the appearance of a specific person. Strata may overlap and thus the meaning of the video at any instance can be flexibly modeled as the union of all strata present. A *stratosphere* [Aquierre-Smith & Pincever 91] has been prototyped to demonstrate the idea of stratification. More recently, the AVIS system (The Advanced Video Information System) [Adali et al 96] developed a formal model similar to the stratification model, but focused more on the efficient organization and data structure of video data. They modeled video sequences as simple entities and developed a linear algorithm to support efficient query processing.

Browsing is another significant task that merits quality computer-based support. The major technological precedent for video browsing is the VCR, with its support for sequential fast-forwarding and reverse play. Browsing a video this way is a matter of skipping frames at an even rate regardless of its content. Therefore there is always the danger that some skipped frames may contain information of greatest interest. One approach to overcome this problem is to employ the segmentation model of detecting cuts, dividing the video into shots, and selecting key frames to represent each shot. The list of key frames is then used as summary of video contents to support effective browsing [Ueda et al 91; Taniguchi et al 95]. [Zhang et al 95] proposed a hierarchical organization of key frames to facilitate browsing. [Hisashi et al 96] presented a method to reduce the browsing space by eliminating repetitive key frames using a combination of chromatic histograms and differences in local average luminance values. A better approach to overcome the browsing problem is to support browsing of video based on contents.

Research on video retrieval in the past has focused on either low level or high level features. The low level techniques have relied on features like color, texture, shape and free text. The resulting system is fully automated but the retrieval effectiveness is limited. The high level techniques model video contents with elaborate concept structure with entities and relations [Prié. et al 98]. They are effective for providing high level queries in a limited domain. However, the setting up of the concept structure is a tedious manual process. Stratification represents a middle ground that models video contents as overlapping strata of concepts. By judiciously choosing the right level of concepts, the extraction of most strata can be automated. It offers high modeling power to permit most complex retrieval and browsing operations to be supported.

This paper describes the design, implementation and testing of the prototype for news video retrieval. The main contributions of this paper include: a) the extension of stratification approach to model audio and structured data, and high level concepts such as categories and events; b) the use of stratification as the basis to support fast content-based browsing; c) the development of an actual prototype which models video contents to demonstrate complex retrieval and browsing operations.

The rest of this paper is organized as follows. Section 2 describes the stratification model. Sections 3 and 4 discuss the design and implementation of the system respectively. The evaluation of the system is presented in Section 5. Finally, section 6 contains the conclusion and discussion of limitations and future work.

2. The Stratification Model

Stratification is a context-based approach to modeling video content as strata. Strata may overlap and thus any frame can have a variable number of strata associated with it. Stratification is therefore a way to model video content as rich multi-layered descriptions that can be parsed by a wide range of applications.

A comparison between the traditional segmentation model and the stratification model is given below.

- In the segmentation model, the basic atomic unit and therefore the granularity of video data is a shot. It is not possible for users to access video contents within the shot boundaries. For stratification, the granularity of information is a frame. It thus permits users the greatest flexibility to access and present video contents.
- The segmentation model imposes authors' intentionality early during the shot creation stage. Thus once the video is segmented, it is difficult to support other users who may need to use the same material for a different purpose. The stratification model views video contents as layers of overlapping strata. Thus users may combine strata to flexibly retrieve video, or easily build additional strata to cater to their specific needs.

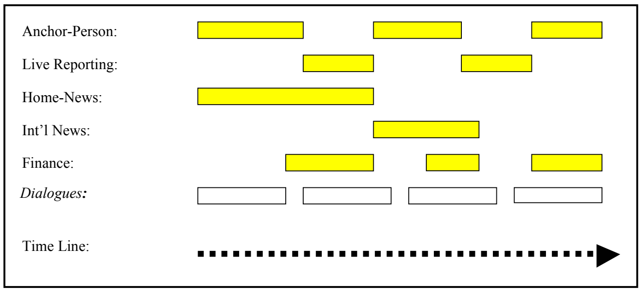


Figure 1: Stratification modeling of a news video

Figure 1 shows an example of the stratification modeling of news video, in which the content is modeled as multiple occurrences of strata. In order to enhance the modeling power of the stratification model, we have extended the model to cover not just single concepts, but also audio dialogue, structured data and high level concepts such as categories and events. Thus, under the extended model, a stratum can be an object such as a person, a category such as the sports news, or audio dialogue. It may occur over many, possibly overlapping time periods.

The strata act as meta-information that provides concise interpretation of video's semantics in terms of multiple layers of concepts. By applying simple union and/or intersection operations, many interesting higher level concepts can be derived from the basic strata. Also, it is possible to use the strata information to support many interesting video manipulation functions such as the content-based fast-forwarding, retrieval, and video summarization. In this paper, we will only present two major applications: video retrieval and fast-forwarding. The research on the other aspects of video management is still going on.

3. System Design

In this section, we describe the design of the Stratification Video Modeling System (SVMS). SVMS is a general-purpose video management system, which supports the whole process of video indexing, retrieval to browsing. We will discuss the details of video representation, query processing and retrieval, and video browsing below.

3.1 Video Representation

3.1a Strata and Frame Sequence

Video can be organized as a set of frames. By following the symbols employed in [Adali et al 96], a frame sequence is defined as a pair [i,j] where $1\leq i \leq j \leq n$. [i,j) represents the set of all frames between i (inclusive) and j (non-inclusive). Thus each strata is mapped to a set of frame sequences to indicate its multiple occurrences in the video. One property that we would like the frame sequence to have is so-called "solid". A set $\underline{X}$ of frame sequences $\{[i_1,j_1) \ldots [i_r,j_r)\}$ is said to be solid if $[i_1, j_1) \subset [i_2, j_2) \subset \ldots \subset [i_r, j_r)$. Here $\subset$ is a partial ordering on the set of all frame sequences as follows: $[i_1, j_1) \subset [i_2, j_2)$ if $i_1 < j_1 < i_2 < j_2$. Intuitively, $[i_1, j_1) \subset [i_2, j_2)$ means that the sequence of frames denoted by $[i_1, j_1)$ precedes the sequence of frames denoted by $[i_2,j_2)$.

We can now briefly define the stratum. As described previously, there are three types of stratum: the first is based on dialogue and the other two are on entity and structured data.

The dialogue stratum D_s of a video stream s is defined as:

$$D_s :== \{(d_{s1},f_1), (d_{s2},f_2), .. (d_{sN},f_N)\}$$

where d_{sk} denotes the paragraph k of audio dialog of video stream s; and f_k denotes the frame sequence that contains dialogue d_{sk}.

The entity stratum E_i is defined as:

$$E_i :== \{e_i, \underline{X}\}$$

where e_i denotes the entity i; $\underline{X}$ is the list of solid frame sequences that e_i appears.

The strata for structured data are defined in the same way as the entity strata.

We now describe each type of stratum in detail.

3.1b The Dialogue Stratum: Paragraphs in a Video Database

Audio dialogue provides rich information about the video and is modeled at the paragraph level. To extract text from audio dialogue, the audio track must be analyzed using speech recognition [Rudnicky et al 94]. There are a lot of commercial and prototype speech recognition systems that could achieve a reasonably high degree of accuracy. Thus in most general applications, the identification of most essential words in a dialogue is not a big problem. We will rely on a commercial IBM product[ViaVoice 98] to automatically identify the majority of sentences on the news video's audio track.

Currently the paragraphs are transcribed manually. Each paragraph is associated with a frame sequence to indicate the start and end frames where the dialogue takes place. An example of a dialogue stratum is shown in Figure 1. From a query, partial match free text retrieval technique [Salton & McGill 83] can be used to identify relevant paragraphs which in turn maps into a list of relevant frame sequences.

3.1c The Entity Stratum

Entities are items that are of interest to the users. How to select an appropriate set of entities to suit an application is an open question that has no uniform answer. Although stratification is rich in data representation, the indexing of entity strata manually is a very tedious task. This is because the same video needs to be parsed multiple times in order to index a good collection of entities. Thus in SVMS, only primitive concepts and well-defined categories are chosen as strata. We are using pseudo-object models together with relevance feedback [Chua & Chu 98] and other learning method to identify and track objects automatically. We are also employing face location [Sung. et al 1998] and tracking methods to identify important faces. We aim to automate as much as possible the process of extracting strata data.

The consideration of what can be automated affects our decisions of what entities should be modeled in SVMS. The various types of entities are discussed below.

- ***Object***: The video objects may be people in the news video such as the Prime Minister of Singapore, Bill Clinton, or anchor person etc. They may also be buildings/places, landmarks, animals or items which are of general interest to the users. Although automatic detection of objects is not possible in the current image processing research, with pre-knowledge of the objects involved, locating and tracking of objects in a video sequence using video analysis techniques is possible using the techniques as described in the preceding paragraph.
- ***Category:*** The semantic structure of news video is quite unique and well structured. Its contents can generally be classified into the categories of: home news, international news, finance, weather, sports and commercial breaks. Among these, the home and international news are rather high-level and require sophisticated analysis for their identification. As it is not possible to automatically identify these high-level categories directly, low level categories such as the shots of anchor person, live reporting, sports, or even reports on money/stock market and weather etc can be used instead. These low-level categories, together with their visual features, audio dialogues, and knowledge of the structure of news video, may be used as the basis to identify high-level categories. This is an area of current research.

- ***Location and Event:*** Location specifies the place where the event took place; and the event entity captures notable occurrences of news items. We are now researching into techniques to identify stories and events using a combination of audio and visual cues together with time based specification of the characteristics of events or stories. These will contribute towards automatic detection of events.

- ***Structured Data:*** In addition to the above, structured data such as the dates, times, and source of video streams etc must be encoded. Such information can be used to support queries based on dates and times such as the “financial news in the last two days”.

Since each entity may occur multiple times within one video stream, each entity is associated with a solid set of frame sequences. When queried about the entities, it is usual to retrieve the whole set of frame sequences that the entities appear. Thus the entity data is organized in an inverted structure with entity names as the inverted index for retrieval. Hence with keys such as Bill Clinton (Object), Indonesia

(Location), or Home News (Category), we are able to retrieve the whole set of frame sequences that these entities appear. Figure 2 shows the data structure for organizing the entity strata.

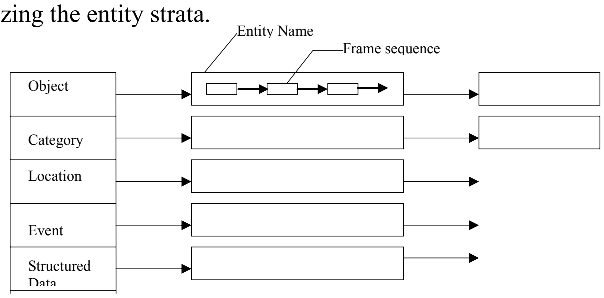


Figure 2: Inverted file structure for entity strata

3.2 *Query Processing and Retrieval*

The semantic of video is modeled as strata of dialogues and entities. We employ logical exact match operations to combine entities to support complex queries; whereas we use free-text partial match technique to locate relevant paragraphs. The combination of both Boolean exact match and IR best match techniques provides a powerful mechanism for querying the database.

3.2a *Free Text Retrieval of Paragraphs*

SVMS uses Vector Space Model [Salton & McGill 83] to support paragraph level free text retrieval. Vector Space model maps each paragraph to a vector space of non-trivial terms of dimension N, where N is number of stems extracted from the paragraphs. The stem terms in each paragraph is indexed using the standard tf*idf term weighting formula. The similarity between the query and a paragraph is simply the distance between them in the vector space. This is computed using the well-known cosine similarity formula as follows

$$Sim(D_i, Q) = \frac{\sum_{k=1}^{t}(d_{ik} \times q_k)}{\sqrt{\sum_{k=1}^{t}(d_{ik})^2 \times \sum_{k=1}^{t}(q_k)^2}}$$

where Sim(D_i,Q) gives the similarity between query **Q** and paragraph $\mathbf{D_i}$. In general, the retrieval returns a list of up to n paragraphs whose similarity values are above a threshold.

3.2b *Logical Operations on Entity Strata*

The exact match technique is used to retrieve entity level strata such as the object, category, location and event. Since the result of the logical AND or OR operation on two solid frame sequences is also a solid frame sequence, the logical operations can be used to construct complex queries based on simple entities.

Figure 3 shows the logical operations on two solid sets of frame sequences. In the *AND* (or *OR*) operation, those frame sequences that *are common to (*or *appear in*) both sets are returned. The exact match is also used to filter video streams based on structured information such as dates, times, broadcast companies or programs etc.

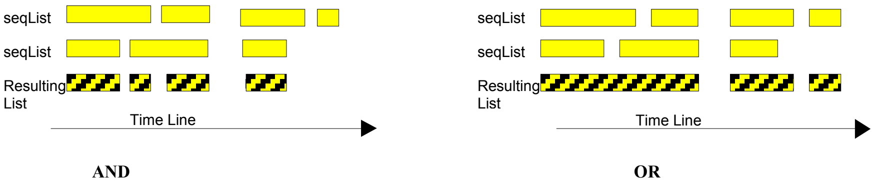


Figure 3: Logical operations on two frame sequences

3.2c Presentation of Query Results

The result of the query is a list of frame sequences. Since the frame sequence can begin or end at any position along the time line of video, a presentation formed by simply concatenating all the frame sequences returned may not be effective. To smoothen out the abrupt effects of simple concatenation, we first extend the frame sequence to match the beginning and end of an audio sentence. We then modify the start frame number by searching for the MPEG GOP (group of picture) where the start frame belongs and starting from the beginning of GOP. Similarly, for each end frame, we search for the GOP that it belongs to and use the ending of GOP as the end frame of our frame sequence. Thus we extend the original frame sequence a bit longer in order to give a smoother presentation with better visual and audio continuity.

The presentation of query result is a major topic of research. In order to have a really meaningful presentation of the query result, other factors such as the semantic and cinematic continuity [Balasz 52] of presentation needs to be considered. This is beyond the scope of this paper.

3.3 Video Browsing

The task of browsing is intimately related to query processing. Furthermore, browsing may serve not only as an aid to the formulation of queries but also as a means for examining the results. Thus an effective browsing tool should base on some level of analysis of the video contents.

As video is modeled as multiple occurrences of simple entities in the stratification model, the meaning of video at any instance is simply the union of strata occurring at that time. This brought us to the idea that fast-forwarding, or more accurately, fast-skipping, can be supported directly using the strata information. For instance, suppose that the user is interested in Bill Clinton, he may choose the entity "Clinton" as the context for browsing. So each time the user clicks the fast-forward button, the video will skip to the next occurrence of Bill Clinton. In this way, the user is able to browse through all the news related to Bill Clinton within a short duration without missing any interesting information. This will be a major improvement over the even-rated fast skipping method used in traditional fast-forwarding.

4. System Implementation

The Stratification Video Modeling System (SVMS) is implemented on the Sun workstation running Solaris 2.0. Figure 4 shows the overall system structure of SVMS. It consists of three modules: *Video Logger*, *Video Query Processor* and *Video Browser*. The *Video Logger* is designed to facilitate the manual indexing of video data. It has a VCR like interface which provides standard functionalities such as play, pause, stop, etc. Two additional buttons are provided to mark the start and end of frame sequences and add in the strata information accordingly. It also provides a stratagraph that gives users an overall view of all the strata information indexed.

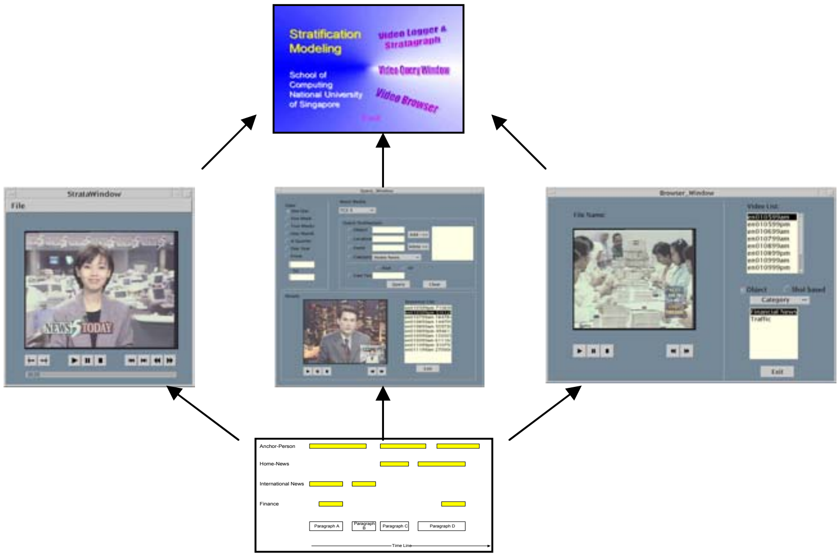


Figure 4: Overall system structure of SVMS

The user interacts with the *Video Query Processor* to perform queries and view the results. The upper part of the screen is the query composer, which supports three types of queries. It supports the free-text queries to locate relevant paragraphs using the IR partial match technique. It permits the users to compose complex entity-level queries by using the AND/OR operator to combine entity specifications. It also supports the use of structured data such as the dates and times to filter video streams. The common interface permits these three types of query specifications to be combined to support complex query like "international news (category) about Bill Clinton (object) in the last two dates (structured data in dates) about 'US-China relations' (free-text description)". The results of the query are presented at the lower part of the screen. A picture in picture display is designed to let the user view the current frame sequence, while previewing the content of the next sequence in the smaller picture display. With the preview sub-window, the user can make the choice of whether to watch or skip the next sequence.

The *Video Browser* facilitates content-based browsing of video. The user can choose any interesting object (indexed in strata) as the context to browse through the entire video database. Whenever the fast-forward button is pressed, the next occurrence of the object is displayed.

5. Evaluation

The system is evaluated using the domain of news video. We utilized one-week of English news from TCS (Television Corporation of Singapore) channel 5 from Jan 5 to Jan 11,1999. For each day, we used both the morning (10 minutes) and evening news (30 minutes). Thus altogether we have close to 280 minutes of news video data. All the news video is stored in MPEG-1 format.

Three types of strata information are indexed: entities, paragraphs and structured data. Four types of entities are identified: object, category, event and location. Only entities involving simple and well-defined concepts are chosen. For this prototype, all entities are extracted manually. However, with the use of simple concepts as entities, it is possible to automate the extraction most strata information. It is the subject of our current research.

Seven types of categories are used. They are the home news, international news, finance, sports, weather, traffic and commercial break. More than 10 objects are identified and each contributes about 5-10 strata to the total database. Each paragraph corresponds to a piece of news. All key words for each paragraph are recorded manually. Altogether, 135 paragraphs are identified.

The following retrieval examples are used to demonstrate the three types of queries supported by SVMS:

- Free-text query: Figure 5 shows the results of issuing the free-text query "currency crisis". The retrieval is based only on the paragraphs indexed. All frame sequences are correctly retrieved except sequence (e).

a) Indonesia currency crisis (05/01/99 am)

b) Indonesia currency crisis (05/01/99 pm)

c) S.K. currency crisis (05/01/99 pm)

d) Asia currency crisis (08/01/99 pm)

e) UK cow disease crisis (10/01/99 pm)

Figure 5: Retrieval results of free-text query “Currency Crisis”

- Query combining both free-text and logical Boolean operation on entity strata: Figure 6 shows the result of the combined query "currency crisis in Indonesia", which combines free-text ('currency crisis') and location entity (Indonesia). All sequences are correctly retrieved.

a) 05/01/99 am

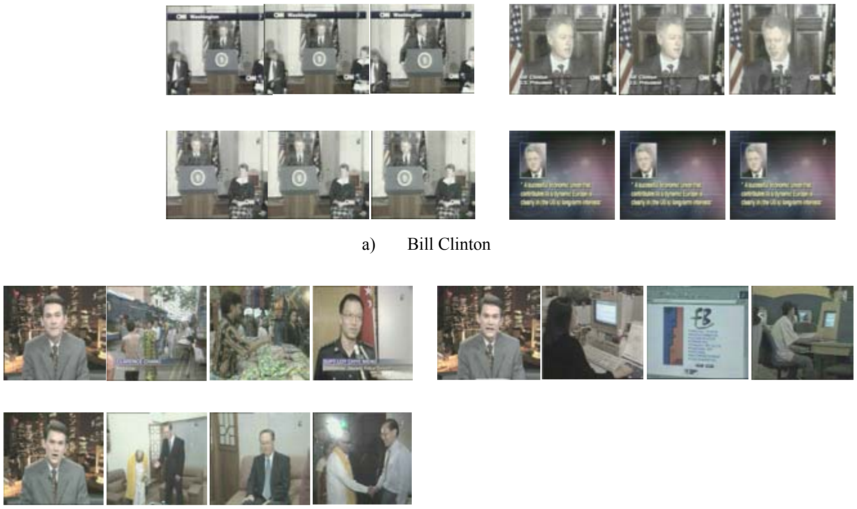
b) 05/01/99 pm

c) 08/01/99 pm

Figure 6: Retrieval result of the query "Currency Crisis in Indonesia"

- Query with additional time constraint: Suppose we specify that we only want news on "currency crisis in Indonesia" from 1st-7thJanuary 1999, then only frame sequences (a) and (b) of Figure 6 will be shown; frame sequence (c) will be excluded.

Figure 7 shows the typical frame sequences encountered by performing the object-based browsing on news of 5th January 1999. Figure 7(a) shows the sequences generated by using the object "Bill Clinton" as the browsing context. The context for browsing for sequences shown in Figure 7(b) is the category "Home News".



a) Bill Clinton

b) Home News

Figure 7: Results of object-based browsing on News on 5th January, 1999

6. Conclusion and Future Work

The aim of this project is to develop a prototype news video modeling system that models the video data using the stratification approach. By using the strata information indexed, we are able to explore advanced video manipulation functionalities such as retrieval and content-based fast-forwarding. The prototype is developed using the news video supplied by the Television Corporation of Singapore. The results of initial testing indicate that these advanced functions can be adequately supported using the stratification approach.

Current research is carried out in the following direction in an attempt to automate the indexing of strata as much as possible.

a) Better techniques for segmenting video sequence into shots and audio stream into sentences. For video sequence, we are developing a new multi-resolution technique[Lin. et al 1999] that could detect a wide range of cut, dissolve and wipe. The shots and sentences provide a basis for identifying stratum boundaries.

b) Techniques for identifying and tracking pseudo objects [Chua & Chu 98] and faces [Sung. et. al 1998] by extending current research in the lab.

c) Techniques to identify stories and events in video sequence. A story is defined as a coherent semantic sequence of news/narratives such as a news item on "Election in Indonesia". A story normally consists of multiple shots with little visual coherence and sentences. An event involves a specified sequence of action that constitute notable occurrences. Examples include riots, war, signing ceremony etc. The stories and events are high level strata.

d) The presentation of retrieval result using both cinematic and semantic rules to ensure visual and audio continuity.

References:

Adali S, Candan KS, Chen SS, Erol K, Subrahmanian VS (1996). "The Advanced Video Information System: data structures and query processing", ACM Multimedia Systems, 4, 172-186.

Aguierre Smith TG and Pincever NC (1991). "Parsing Movies in Context", USENIX-summer'91, 157-168.

Balazs B (1952). "Theory of the Film". London, Dennis Dobson Ltd.

Chua TS & Chu CX (1998). "Color-based Pseudo Object Model for Image Retrieval with Relevance Feedback". Proceedings of 1st International Conference on advanced Multimedia Content Processing, Nov, Osaka, Japan, 148-162.

Chua TS & Ruan LQ (1995). "A Video Retrieval and Sequencing System". ACM Trans of Information Systems, 13(4): 373-407, Oct.

Hisashi A, Shigeyoshi S & Osamu H (1996). "A Shot Classification Method of Selecting Effective Key-frames for Video Browsing". ACM Multimedia ‘96, ACM Press, 1-10

Y. Lin, T. S. Chua, M. S. Kankanhalli, (1999). "Temporal Multiresolution Analysis for Video Segmentation". Technical Report, School of Computing, NUS, 1999

MPEG (1993). "Information Technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s", Moving Picture Expert Group Committee, ISO/IEC 11172-1,2,3,4.

Yannick Prié, Alain Mille, and Jean-Marie Pinon (1998). "AI-STRATA: A User-centered Model for Content-based Description and Retrieval of Audiovisual Sequences". Proceedings of 1st International Conference on advanced Multimedia Content Processing, Nov, Osaka, Japan, 333-348

Rubin B & Davenport G (1989). "Structured Content Modeling For Cinematic Information", SIGCHI Bulletin, 21(2), 78-79.

Rudnicky AI, Lee KP & Hauptmann AG (1994). "Survey of current speech technology". Communications of the ACM, 37(3): 52-57.

Salton G & McGill M (1983). "Introduction to Modern Information Retrieval", McGraw-Hill, New York.

Sung, Kah Kay and Poggio, Tomaso (1998). "Example-based Learning for View based Human Face Detection". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, 39-51.

Taniguchi Y, Akutsu A, Tonomura Y & Hamada H (1995). "An Intuitive and Efficient Access Interface to Real-time Incoming Video Based On Automatic Indexing", ACM Multimedia 95, ACM Press

Ueda H, Miyatake T & Yoshizawa S (1991). "Impact: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System", Proc. ACM Multimedia 91, ACM Press, 343-350

ViaVoice (1998). IBM's speech recognition technology. http://www.software.ibm.com/speech/

Zhang HJ, Smoliar SW & Wu JH (1995). "Content-Based Video Browsing Tools". Proc. SPIE Multimedia Computing and Networking, 2417: 389-398.