社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－ＡＶＭ－43 (23)
2003／12／19

Hierarchical Video Modeling for Indexing and Retrieval Based on MPEG-7

SHEN Jinhong Seiya MIYAZAKI Terumasa AOKI Hiroshi YASUDA

School of Engineering, The University of Tokyo 4-6-1 Komaba, Mekuro-ku, Tokyo, 153-8904 Japan
E-mail: {j-shen, seiya, aoki, yasuda}@mpeg.rcast.u-tokyo.ac.jp

Abstract To reuse abundant video assets upon demand, we need advanced techniques for flexible video indexing and retrieval. Our video model is designed for a desktop movie making system DMP (Digital Movie Producer) we are implementing which can interpret a verbal screenplay into a relevant motion picture automatically with various visual effects like real image, 3D animation, or their composition. This work proposes a video retrieval system with the hierarchical structure based on suitable semantic annotation combining low-level video features and high-level concepts. It contains three sides of works: extracting semantics automatically, storing these semantics in MPEG-7 metadata format, and retrieving the semantic video content based upon the metadata.

Keyword Video Retrieval, Video Modeling, Video Index, Content-based Retrieval, MPEG-7, Knowledge-based

1. Introduction

In the large body of knowledge that surrounds learning styles, humans learn 83% through sight, 11% through hearing. A low-cost easy-to-use moviemaker system has good entertainment and education markets. We are implementing such a technique DMP (Digital Movie Producer) by which nonprofessional can make and deliver his own movie easily [1, 2]. DMP aims to interpret a verbal screenplay into a relevant motion picture automatically with various visual effects like real image, 3D animation, or their composition, where real images are extracted from digital video library (Figure 1).

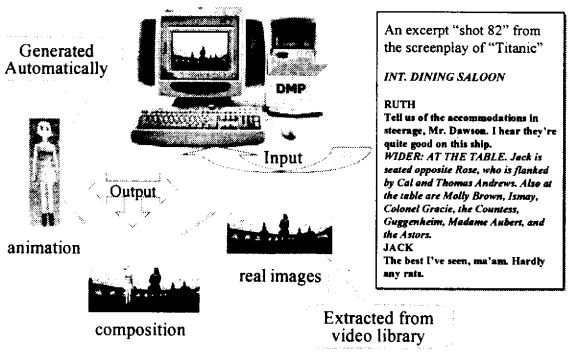

Figure 1. DMP System Architecture

Video is the sort of multimedia full of vast information. It enables us to absorb information the most interestingly, conveniently and effectively compared with pure text and image. To reuse the video, the whole processing for retrieval involves content analysis and feature extraction, content modeling, indexing and querying [3]. In this paper, we mainly introduce a video data model and an ontology that possesses sufficient semantics effectively describing its abundant features for both information extract and retrieval to be used by virtual film director of DMP. The virtual director is responsible for the visual aspect of screenplay dependent on knowledge of plot structure in knowledge base and screenplay.

2. Video Search Task from the Point of View of Director

A film is made up of shots arranged in sequence. The virtual director establishes a point of view on the action that helps to determine the selection of shots and camerawork through rule-based planning, timing out every shot and important camera move. He first makes high-level shooting plan such as "track one's face" for each event based on his directorial expertise, then gives commands about shot types and shot sequence, at last calculates the parameters of camera position, orientation, and movement to satisfy the these commands. Generally his knowledge can be classified into three types:

1. World knowledge: general information that supports commonsense reasoning such as New Year's Day is January 1rst;
2. Cinematic knowledge: mise-en-scène (what to shoot), cinematography (how to shoot), montage (how to present the shots), and sound related to image;
3. Domain knowledge: special information about domains such as sports (soccer) and dance (ballet).

Extraction of domain knowledge means to make up the gap between the data models of these types of knowledge and database schema. Raw video naturally has a hierarchy of units from base level of individual frames to higher

levels of *segments* such as *shots*, *scenes*, and *episodes*. We defined shot as the single uninterrupted operation of the camera that results in a continuous action. A *scene* contains a group of shots which depict an event in the story and occur in one place. *Event* is an important primitive action unit in camera planning procedure such as "a private conversation between two characters" (two-talk). A series of related scenes form an *episode*. An important task in analyzing video content is to detect segment boundaries [4].

Second, director gives commands for the dramatic structure, pace, and directional flow *elements* of the sounds and visual images to visualize the event. Composition, the location of characters, lighting styles, depth of field and camera angle are all determinant factors in the formulation of the visual information. He will use content information obtained in audio/video like space, time, weather, characters, objects, character actions, object actions, relative position, screen position, cinematography, dialog, music, laugher etc [5].

Third, varied video information needs to be organized in a structured fashion – *video data model* to present various types of multimedia information [6]. In DMP we should employ suitable data model based on MPEG-7 in order to provide more effective and efficient video retrieval. MPEG-7 standard has been used to encode video data by DMP for MPEG-7 is mainly intended for content identification purposes while other coding formats such as MPEG-2, 4 are mainly intended for content reproduction purposes.

3. The Related State of Art of Video Retrieval

Current image and video retrieval systems are the results of combining research from various fields: computer vision, multimedia database, artificial intelligence, natural language understanding, pattern recognition as well as digital signal processing and statistics.

3.1. Classes of Video Information

From a point of view of user's needs, video information can be grouped as *bibliographic* (e.g. video title), *structural* (composition of temporal segments: e.g. shot) and *content information* (information which can be seen an understood).

From the point of view of data analysis, video surrogates can be classed under the headings *raw video features* (e.g. file size), *physical features* (spatio-temporal distribution of pixels: e.g. color) and *semantic features* (high-level concept: e.g. object). Combination of the above two types of classification is showed in table 1.

3.2. Content-based Video Retrieval (CBVR) System

There two main categories of the video retrieval approaches. (1) *Anotation-based approach* uses keyword, attribute or free-text to present high-level concepts of video content usually by manual annotation. The procedure of annotation is tedious and consuming. It is difficult to annotate by automatic way because there is gap between low-level feature and high-level concepts. (2) *Content-based video retrieval approach* depends on the understanding of the content of multimedia documents and of their components. Query like "find red ball moving from left of the frame to right" relates to primitive level of video content (color, texture, shape, motion); query like "a plane taking off" relates to high-level content (named types of action), query like "an video depicting suffering" relates to higher abstract level (emotion). How to understand the contents of video?

To date, several research (Photobook, VisualSEEk) and commercial (QBIC, Virage) systems provide automatic indexing and querying based on visual features such as color and texture. While low-level visual content can be extracted automatically, extracting semantic video features event automatically such as is still difficult, and it is usually domain dependent such as on sports [7, 8].

3.3. Data Modeling

Data model deals with the issue of representing the content of all media objects present in video database. In detail the work is to design high/low-level models of raw video concerned with various operations including media selection, insertion, editing, indexing, browsing, querying, and retrieval. The temporal nature of video data and requires special query and retrieval functions different

Types	Raw video features	Physical features	Semantic features
Bibliographic information	Coding format, frame rate, video duration, etc		People, place, etc
Structural information		Spatio-temporal structure	Shot, scene, camerawork
Content information		Color, texture, shape, sketch, motion,	Object (face), action, event, episode, emotion, etc

Table 1. Classed of Video Information

from image's. A video data model intended for movies was given by Corridoni et al in 1996. The model grounded in film theory facilitates the querying and browsing of digital video data based on combination of content description as well as technical film feature (camerawork, editing) and structure (shot described in terms of objects and actions it contains).

4. Architecture of Our DMPVR Subsystem

Systems that combine visual features, sound, text as well as structured descriptions will get powerful retrieval. We will use textual information such as closed captions whenever available for video indexing.

4.1 Video Segmentation

Shot change detection may be realized by *direct pixel comparison* or a more robust method *histogram comparison*. A shot change is detected if a significant percentage of pixels differ or if the histograms of two consequent frames differ significantly. (Camera operations such as zooming, tilting, and panning will make it difficult to detect shot changes.)

After the video is divided into different shots by using one or more of the above techniques, the shots are classified based on the models (e.g. weather forecast, news).

4.2 Automated Annotation

To detect and track *objects*, a typical strategy is to initially segment regions based on color and texture information. After the initial segmentation, regions with similar motion vectors can be merged subject to certain constraints such as adjacency. Human *faces* can be detected by using human skin color and DCT transform coefficients in MPEG and broad shape information. It is possible to recognize certain facial expressions and gestures using models of face or hand movements. Particular movements such as entering/exiting a scene and positioning objects using motion vectors are able to detect.

Another strategy is domain dependent way by priori knowledge-based reasoning off-line demonstrated in figure 2.

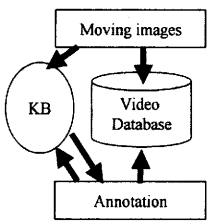

Figure 2. Knowledge-based Semantics Extracting

4.3 Content Modeling, Indexing, and Retrieval

From a point of view of filmmaker, generating motion pictures employing cinematic techniques involves information of the real world, shot sequences composition and how to shoot them. This semantic information like event can be extracted directly from audio-visual features (coming from visual contents, sound, integral and external text) in some domains by knowledge-based approach. Our video content modeling scheme is showed as the following (table 2).

Model Categories	Explanation
World Model	Foreground: e.g. prop, character, hair color, consume, Background: e.g. place, weather
Cinematic Model	Camerawork: e.g. pan, long shot, Editing: shot, scene,
Domain Model	e.g. spots game (foul rule base, team structure, temporal model)

Table 2. Video Content Modeling

where in *world model* character, prop (colour of the ball or structure of basket in basketball game); in *cinematic model* zoom in/out, pan left/right, camera location; in *domain model* object-oriented hierarchical model (e.g. quarter1,break1, quarter2, break2, ... end in basketball game), rule base (e.g. game rules), temporal model (e.g. temporal structure of game) have the possibility to be annotated automatically at present.

Our query scheme approach depends on media features, visual features, and semantic features of the above video data model, explained by the following table 3.

Multimodal Query	Retrieval Items
Query by example	Visual features
Query by text (Keywords and free-text)	Cinematic structure Semantic content (of annotated video)
Query by standard query language	Semantic content (of un-annotated video)

Table 3. Video Query Scheme

We use *text query*, *image/video example query* but no sketch query.

4.4 XML-based Ontology

MPEG-7 (DSs, Ds, DDL based on XML) standardizes the information exchange of descriptive information. We use its low-level and high-level descriptive metadata for video data modeling and retrieval. But only MPEG-7 is not suitable enough to serve as a multimedia data model, for its aim was not taking into different purposes. In DMP, XML tags are supported by our DMPML (e.g. figure 3, 4).

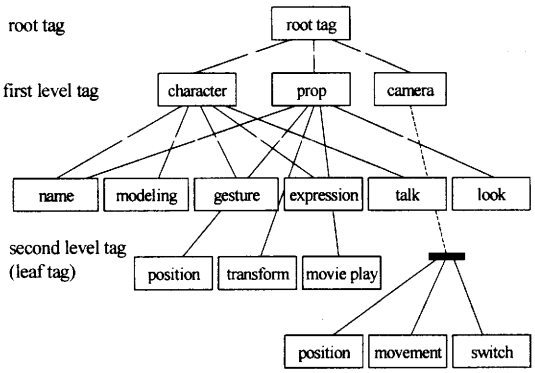

Figure 3. An Example DMPML File

```
An DMP DTD Example extracted from file.txt
<!ELEMENT characters (character *)>
<!ELEMENT characters (name, modeling, gesture, expression,
      talk, look)>

An DMPML Example extracted from file.XML
<?XML version="1.0"?>
 <DMPML>
      <characters>
         <character>
                  <name>Jack </name>
                  <modeling> sportsman </modeling>
                  <gesture> stand </gesture>
                  <expression> smile </expression>
                  <talk> I am fine</talk>
                  <look> blink </look>
         </character>
      </characters>
</DMPML>
```

Figure 4. Family Tree of the above Example

4.5 Other Approach of Automation

Stochastic method that often use automatic learning ability to extract event like explosion are being considered like Hidden Markov Models (HMMS) or Dynamic Bayesian Networks.

5. Conclusion

We describe a multi-category video modeling and multi-modal query mechanism constructed from the perspective of filmmaker for the motion picture generation technique DMP we are implementing. The video indexing subsystem is operated based on MPEG-7 to take advantage of its metadata for the effective retrieval of video data. Media features (e.g. coding format), visual features, and semantic features are already labeled with video or can be obtained by various corresponding algorithms. Our research question lies in how to organize these data for effective and efficient query applied for the use in DMP system.

Query and transaction models of video database systems differ from those of the traditional database systems. With the advancement of techniques on computer vision and multimedia database, video retrieval systems developed from *traditional text-based* video indexing annotated manually (using keyword, attribute, free-text to present high-level concept), *content-based* video indexing exploiting the technique of signal processing (focusing mainly on extracted low-level visual features: color, shape, texture, motion), to current *semantics-based* video indexing by semantic annotation exploiting the techniques of Artificial Intelligence (high-level semantic features: object, event; and higher-level semantic features: emotion). But it is still not easy to be annotated automatically, only realized in some domains such as sports (basketball) and dance (ballet).

Reference

[1] SHEN Jinhong, Seiya MIYAZAKI, Terumasa AOKI, Hiroshi YASUDA, "Filmmaking Production System with Rule-based Reasoning", Image and Vision Computing New Zealand (IVCNZ 2003), Palmerston North, New Zealand, Nov. 26-28, 2003

[2] SHEN Jinhong, Seiya MIYAZAKI, Terumasa AOKI, Hiroshi YASUDA, "A Prototype of Cinematic Rule-based Reasoning and Its Application", CCCT'03, The 9th International Conference on Information Systems Analysis and Synthesis: ISAS '03, July 31- Aug. 2nd, Orlando, Florida, USA, 2003

[3] Aslandogan, Y. and Yu, C., Techniques and Systems for Image and Video Retrieval, IEEE Transactions on Knowledge and Data Engineering, pp.56-63, 11(1), 1999.

[4] G. Ahanger and T.D.C. Little, "Automatic Digital Video Production Concepts," Handbook on Internet and Multimedia Systems and Applications, CRC Press, Boca Raton, FL., December 1998.

[5] M. F. McTear, Spoken dialogue technology: enabling the conversational user interface, ACM Computing Surveys, vol. 34, pp. 90 - 169, 2002.

[6] M. Petkovic, "Content-based Video Retrieval", VII. Conference on Extending Database Technology (EDBT), Ph.D. Workshop, Konstanz, Germany, March 2000.

[7] H.J. Zhang, John Y. A. Wang, and Yucel Altunbasak. "Content-based video retrieval and compression: A unified solution", In Proc. IEEE Int. Conf. on Image Proc., 1997.

[8] Salwa, Video Annotation: the role of specialist text. PhD Dissertation, Dept. of Computing, University of Surrey, 1999