2004－CVIM－145　(14)
2004／9／10

Practical Video and Animation Retrieval in E-moviemaking System

Jinhong Shen　　Terumasa Aoki　　Hiroshi Yasuda

Abstract In this paper, we mainly introduce how to apply current advanced techniques on implementing video annotation-metadata-based and content-based retrieval system EMMVR for reusing video clips and computer-generated animation extracted from video library, where EMM (Electronic-MovieMaker) is a software tool we are developing that can interpret the textual screenplay into the movie with the visual effects of 3D animation and real images, and their simple composition. A suitable multi-category (hierarchical and stratified) video modeling and multi-modal query by text and example mechanism with semantic video indexing using visual features, non-visual features, and sound cues are constructed based on MPEG-7 from the perspective of virtual film director by taking advantage of filmmaking ontology.

1. Introduction

Among applications of technologies for computerized management, *Database Systems* are developed for numeric and format texts, *Information Retrieval* is available for users to access unformatted texts, with the ascendant of large collections of video documents, *video storage and retrieval* have become a challenge for us to take up. If video media data are analog, archives related to the cassettes of video may be computer management. But it is impossible to control analog media data and its digital archives in an integrated environment digitally. Today's multimedia technology makes digital video storing and processing possible, further more enables video to be stored as virtual video documents. Database techniques are evolving towards various multimedia systems such as those of sharing, reuse, and retrieval for video information.

This paper describes a specific application of video retrieval which will be combined in a software tool named EMM (Electronic-MovieMaker) we are implementing in order to interpret textual screenplay into sound motion picture with the visual effects of 3D animation and real images, and their simple composition [1], [2]. The real images may be digital video or film. Strictly speaking, *DVD videos* are not as the same as digital films because they can utilize various lossy formats related to MPEG but digital film is lossless - uncompressed or sometimes compressed losslessly. There are various different digital video types. MPEG, Quicktime, Avi, Shockwave, Flash and Animated GIF's are current common formats. To view most of these files types viewers need to download a Plug-in or Movie Viewer.

In our design, a suitable multi-category (hierarchical and stratified) video modeling and multi-modal query by text and example mechanism with semantic video indexing using visual features, non-visual features, and sound cues are constructed based on MPEG-7 from the perspective of virtual film director by taking advantage of filmmaking ontology. The media database of digital video contains digital film, and the media retrieval concerns animation as well.

The next section will have an overview of the related works of other researchers. In Section three, theory basis and requirements in the design of video/animation retrieval are introduced. Section four first outlines the architecture of EMMVR system, then expounds how to realize the automation of video retrieval and animation by using content-based approaches from a film director's point of view. Finally, I will have a discussion about my present research and future work.

2. Related Video Retrieval Researches

There are many classes of video applications: *interactive video*, *video-on-demand*, *stock-shot*, *video edition*, etc. For indexing and retrieval of video data,

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

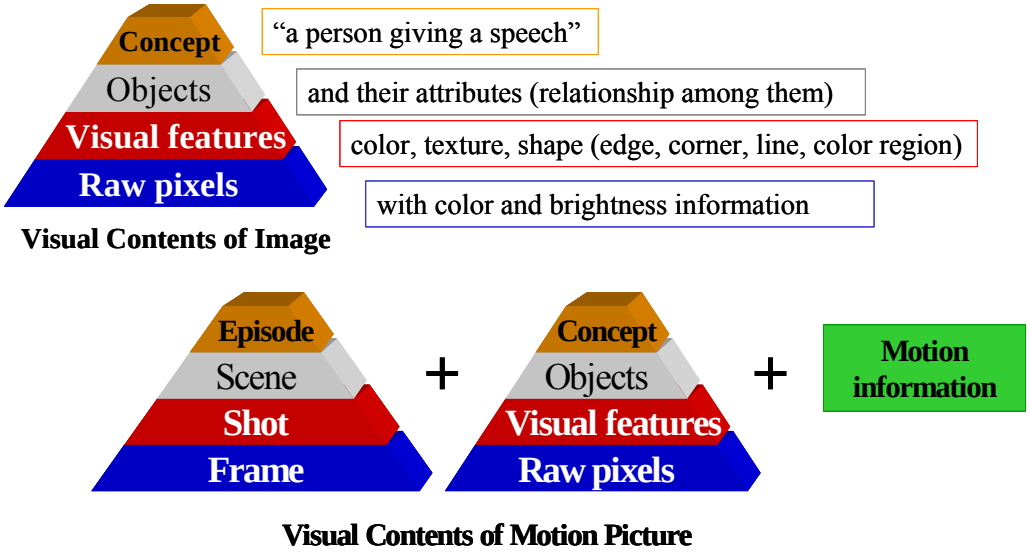

Figure 1. Visual Contents of Image/Video

a variety of methods have been proposed such as *text-based*, *metadata-based, content-based,* and *integrated* approaches.

Metadata-based Video Retrieval

Besides video media data themselves, metadata also give important video information. Typically metadata contains the video specific data such as video name, date of production, length of video, etc.

Text-based Video Retrieval

This traditional method uses keyword, attribute or free-text to present high-level concepts of video contents that are usually annotated manually. Most frequently used image/video retrieval systems are oriented around text searches, for example, www.google.com where textual annotation was already performed. But the procedure of annotation is so tedious and consuming, and there is no standard for video depiction. These drawbacks made researchers to explore content-based way for video indexing and retrieval.

Content-based Video Retrieval (CBVR)

Content-based video retrieval concerns the techniques that capture the spatio-temporal distribution of pixels. In this case *content* refers to the properties of image/video data, rather than the meanings viewers percept directly so that it is possible to annotate automatically by computed way. However it is not always successful because there is gap between low-level feature and high-level concepts. CBVR is a database perspective method depending on the understanding of the content of multimedia documents and of their components.

Several researches (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 automatically such as event is still difficult, and it is usually domain (e.g. sports, dance) dependent [3], [4].

In our considerations, whereas the user has inherent information need expressed in semantics of query, or high-level concepts, the system operates according to the low-level features. That is to say either the user has to make the semantic-content translation or has to find a suitable video clip (or keyframe) to represent the query. We proposed annotation-metadata-based and content-based retrieval system EMMVR (Electrical MovieMaker Video Retrieval) for reusing video clips and computer-generated animation extracted from video repository.

3. Visual Information for EMMVR

3.1 Common Visual Contents

From the data analysis perspective, 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). These visual contents are grouped in hierarchical layers as showed in Fig. 1.

Query like "find red ball moving from left of the frame to right" relates to primitive level of video contents (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). Building semantics from raw video data becomes the main problem of content-based video retrieval.

3.2 Video Structure

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*. *Shot* is defined as the single uninterrupted operation of the camera that results in a continuous action. A film/video is made up of shots arranged in sequence. A *scene* contains a group of shots that depict an event in the story and occur in one place. Concept *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 [5].

3.3 Ontology Coded in XML

In EMM a virtual director gives commands for the dramatic structure, pace, and directional flow *elements* of the sounds and visual images to visualize the event. 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 [6]. Composition, the location of characters, lighting styles, depth of field and camera angle are all determinant factors in the formulation of the visual information.

Ontology can be seen as a conceptual map where the links between individual pieces of filmmaking knowledge are delineated. In the ontology tree of figure 2, those dark squares indicate the main contents should be extracted from video in order to reuse the video for digital moviemaking.

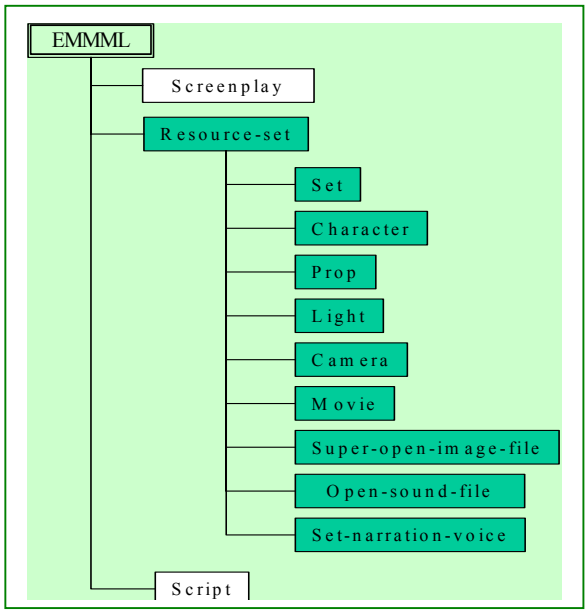

Figure 2. Visual Contents Tree in EMM System

Under the type of <Resource-set>, those subtypes of background <Set>, character <Character>, object <Prop>, lighting information <Light>, camerawork <Camera>, moving picture <Movie>, static image <Super-open-image-file>, music and effects <Open-sound-file>, dialogue or talk <Set-narration-voice> will be utilized to describe the features and contents of videos in library.

MPEG-7 standard has been used to encode video data because 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. For MPEG-7 (DSs, Ds, DDL based on XML) standardizes the information exchange of descriptive information [7] [8], we use its low-level and high-level descriptive metadata for video data modeling and retrieval. But only MPEG-7 is not completely suitable enough to serve as a multimedia data model, for its aim was not taking into different purposes. In EMM, XML tags are supported by our EMMML (XML of EMM). An example EMM XML description of character element and its features like age hair clothes is showed in figure 3.

But based on current technology, not all of the information can be extracted and annotated automatically.

```
<RESOURCE-SET>
 <Character name="Father" cid="F011/123456789ABD"
            type="urn:u-tokyo:dmp:cs:v0.5:Object:3DModel"
            href="http://foo.tv/Father.jar">
  <Feature type="Format" value="TVML Character" />
  <Feature type="Type" value="Human" />
  <Feature type="Gender" value="Male" />
  <Feature type="Age" value="Middle" />
  <Feature type="Voice:Style" value="Deep" />
  <Feature type="Voice:Language" value="English" />
  <Feature type="Hair:Style" value="Casual" />
  <Feature type="Hair:Color" value="Black" />
  <Feature type="Hair:Length" value="Short" />
  <Feature type="Skin:Color" value="Yellow" />
  <Feature type="Eye:Color" value="Brown" />
  <Feature type="Glasses:Style" value="Two Point" />
  <Feature type="Clothes:Shirts:Style" value="Open Neck" />
  <Feature type="Clothes:Shirts:Sleeve" value="Short" />
  <Feature type="Clothes:Shirts:Color" value="Striped Blue" />
  <Feature type="Clothes:Trouser:Style" value="Jeans" />
  <Feature type="Clothes:Trouser:Color" value="Blue" />
  <Feature type="Clothes:Trouser:Length" value="Long" />
  <Feature type="Action" value="Walk" />
  <Feature type="Action" value="Talk" />
 </Character>
</RESOURCE-SET>
```

Figure 3. Contents Description in XML Language

4. EMMVR Design

Giving a description of EMMVR (EMM Virtual Retrieval, see figure 4) in one sentence, it is a subsystem of EMM with a suitable multi-category video modeling and multi-modal query mechanism with multi-modal video indexing from film director's viewpoint. EMMVR focuses on design multi-modal video (and animation) indexing.

4.1 Video and Animation Analysis

The subsection reviews the internal representation for deriving a representation of a multimedia document of video and animation automatically.

Video/Animation Parsing

Video partitioning can operate at four levels of granularity: video-level, scene-level, shot-level frame level: A *scene* is a set of contiguous shots having a common semantic significance. The partitioning of the video into shots uses the temporal information, but generally does not refer to any semantic analysis. Types of shot boundaries like *cut, wipe*, and *dissolve* can be recognized. At frame level, there is little or no temporal analysis.

Feature and Event Extraction

Feature extraction (Table 1) is distinguished into generic feature extraction and description feature recognition supervised by heuristics or training. These operations rely on the analysis of the Human Visual System (HVS), ranging from simple statistics to elaborated model-based filtering techniques. The event information can be extracted directly from audio-visual features (coming from visual contents, sound, integral and external text) in some domains by knowledge-based approach.

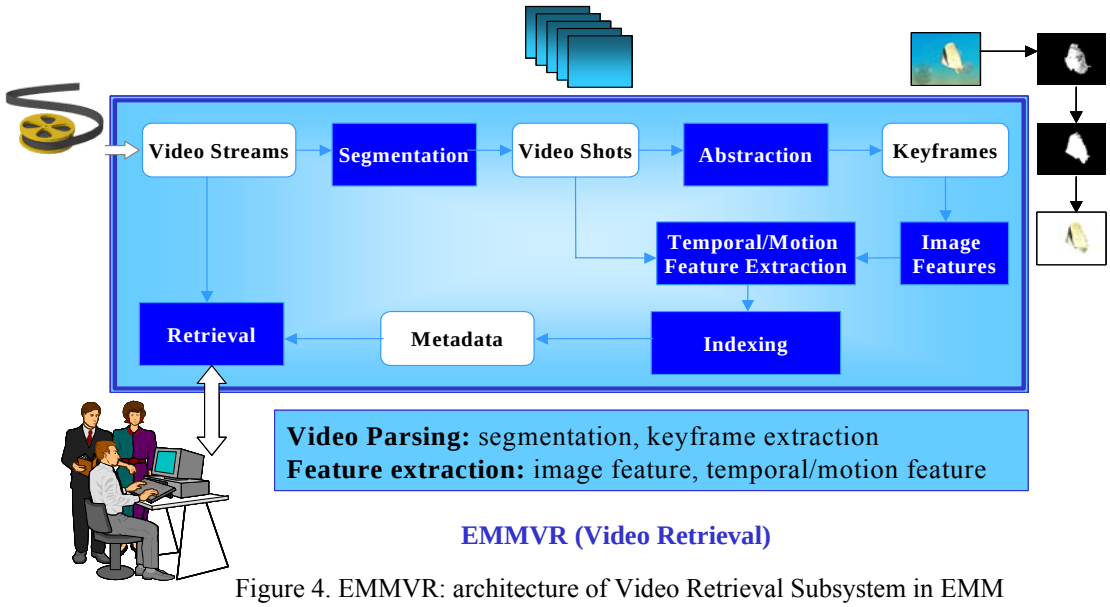

Figure 4. EMMVR: architecture of Video Retrieval Subsystem in EMM

Approaches	Tasks
Automatic annotation	1. Segment (vs. montage): Scene → Shot → Keyframe. 2. Feature extraction: (vs. mise-en-scène) Generic visual spatial features: Color and texture Semantic features: (vs. mise-en-scène) – Face, character, prop and set in specific domain; – Motion feature Camera motion like pan; Object motion – Audio feature (vs. sound edition) Sound like music, dialogue. 3. Event extraction: (vs. mise-en-scène & sound edition) e.g., sport type.
Computer aided annotation	User provides indices through interface of the software detector.

Table 1. Video Information Recognition Tasks

For animation, low-level image features can be recognized based on Attribute Relation Graphs (ARG). More complex objects are derived through the image analysis.

The above tasks are inversive procedures to filmmaking techniques involving the four aspects from film theory:

- *mise-en-scène* (*what to shoot*) which involves setting, lighting, figures,
- *cinematograph* (*how to shoot it*) which involves camerawork – camera angle, camera movement and camera distance,
- *montage* (how to present the shots), e.g., fade in/out, parallel editing,
- *sound edition* (how to present the sounds), e.g., dialog, music, background sound.

Automated annotation is realized if possible by content-based approach. But *fully automatic semantic annotation* is still impossible with current VR technology. For the contents that cannot be annotated automatically, *computer aided content indexing* (Fig 5) may be chosen as a feasible way for complement.

4.2 Video Querying

Visual content may be conveyed in both narrative (language) and image. The detailed explanations of *Multi-modal query mechanism* are showed in table 2.

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

Table 2. Query Mechanism

Varied video information needs to be organized in *video data model* – a structured fashion to present various types of multimedia information, constructed based on MPEG-7. By taking advantage of ontology as mentioned in the above section, it is possible to facilitate *multi-category video modeling*. Systems that combine visual features, sound, text as well as

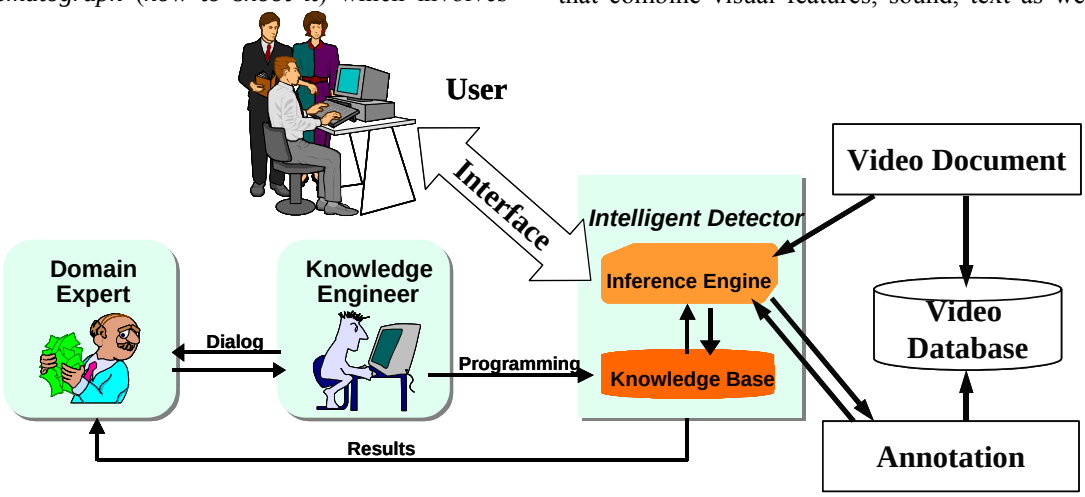

Figure 5. Computer-aided Annotation

structured descriptions can get powerful retrieval. We will use textual information (such as closed captions) whenever available for video indexing.

5. Conclusion

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). We combine annotation-based and content-based retrieval together in our video/animation retrieval system EMMVR design.

This paper describes a multi-category video modeling and multi-modal query mechanism constructed from the perspective of filmmaker for the motion picture generation technique EMM 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 EMM system.

References

[1] Shen Jinhong, Seiya Miyazaki, “Terumasa Aoki, Hiroshi Yasuda, The Framework of an Automatic Digital Movie Producer”. 2002 AVM Conference of IEICE, IEICE Technical Report, 102, 517, Nagoya, Japan, Dec 2002, p.p. 15-18.

[2] Shen, Jinhong; Miyazaki, Seiya; Aoki, Terumasa; Yasuda, Hiroshi, “A Prototype of Cinematic Rule-based Reasoning and Its Application”. The 9th International Conference on Information Systems Analysis and Synthesis: ISAS '03 (CCCT2003), VI, Florida, USA, Aug 2003, p.p. 60-365.

[3] 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.

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

[5] 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.

[6] M. F. McTear, “Spoken dialogue technology: enabling the conversational user interface”, ACM Computing Surveys, vol. 34, 2002, p.p. 90 – 169.

[7] Smith, Manjunath, Day, ICCE 2001 MPEG-7 Tutorial Session, 6/17/2001

[8] MPEG 7 Main Page http://www.darmstadt. gmd.de/mobile/MPEG7/