2004－CG－116　(10)
2004／8／19

EMM Software System: Electronic Movie Making from Screenplay

Jinhong Shen　　Terumasa Aoki　　Hiroshi Yasuda

Abstract This paper describes a software system EMM (Electronic Moviemaker) designed to visualize user's inputted screenplay words by sound motion picture with effects of real image, three-dimension animation, or their composition. A virtual director achieves user's intentions by knowledge-based (KB) approach through setting a scene, determining the corresponding shot types and shot sequence, and planning virtual camerawork dependent on the cinematic expertise stored in a domain knowledge base, where real images are extracted from digital video by applying advanced content-based retrieval techniques, animation generation is automated by interpreting textual screenplay into TVML language to show on its player.

1. Introduction

Our society is becoming visual mediated society of images for advanced broadband network techniques and television culture. If a persons learns how to make his presentation and deliver it uncomplicatedly, the switch will be made from passive watching to active using by visual communication. A low-cost easy-to-use electronic moviemaker has good entertainment and education marketplace.

After analyzing the feasibility of realization, we came to the conclusion that it is reasonable to automatically visualize a verbal screenplay using relevant sound motion picture with visual effects like real image, 3D animation, or their composition, where real images are extracted from digital video (movie, animation, TV programs, etc.) library [1], [2]. The production system EMM (Electronic Moviemaker) we are implementing can understand user's input screenplay through a parser then automatically interprets it into a relevant sound motion picture under the direction of a virtual director in place of a human one dependent on filmmaking knowledge base (KB), where real images are extracted from digital video by applying advanced content-based retrieval techniques, animation generation is automated by interpreting textual screenplay into Japanese NHK's TVML Language to show on TVML Player 1.2. For those nonprogrammers, the cinematic knowledge-based environment instead of programming will save them time and labor greatly.

The next section will have an overview of the related works of other researchers. In Section three, the design of screenplay user interface is introduced. Shot sequence is generated from screenplay that describes abstract relationship between objects (such as *two talk*) or concrete actions of characters (such as *stand*) in various shots (such as *close up*). Section four expounds how to realize the automation of video retrieval and animation generation in software system EMM by using AI approaches from a film director's point of view. Finally, I will have a discussion about my present work.

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

2. Related Work

Works on interpreting text-based input into dynamic visualized presentations are in progress like *Virtual Director* in [3] and *Mario* in [4]. Virtual Director aimed to visualize simple scenario in virtual scene and animation. Mario focused on automatic camera control to create 3D animation from annotated screenplay. They were both designed through KB approach but not systems for home movie making usage. Other methodologies employing AI for computer animation have been put forward. In [5], [6], domain knowledge base was applied in automatically generating animation focusing on camera shot design while in [7] animation creation focused on human gesture. Cognitive modeling for intelligent agent was employed by John Funge et al to solve the same cinematic problem [8].

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). 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 automatically such as event is still difficult, and it is usually domain dependent on such as sports [9], [10].

Some works on translating verbal presentations into visualized presentations are in progress. At & T' is making a system named WordsEye [11] for automatically converting text into representative 3D static scenes. As they said, "Natural language is an easy and effective medium for describing visual ideas and mental imaginary." However, fully capturing the semantic content of language in movies is infeasible because linguistic descriptions tend to be at a high level of abstraction and there will be a certain amount of unpredictability in translating the script into the visual effects.

We chose screenplay as input because it is a formal language for filmmaking that implys the lots of rules of film that are almost invisible by audience. By Artificial Intelligent (AI) approach, the digital filmmaking procedure can be further automated dependent on cinematic knowledge base. A virtual director achieves user's intentions by knowledge-based approach through setting a scene, determining the corresponding shot types and shot sequence, and planning virtual camerawork dependent on the cinematic expertise stored in a domain KB.

3. Screenplay Formats

When designing the form of screenplay, the first feature to consider is *common user access*: it should be easy-to-learn and easy-to-use for non-professionals and non-artist users such as school-age children. On the other hand, 3D motion picture is far more difficult to be realized than 3D static scene, decided by synthetic techniques involving the fields of Linguistics, Artificial Intelligence, Computation, and Computer Vision so that the screenplay design is also based on current technology. The two kinds of screenplay formats were designed by carefully analyzing the correspondence between words and pictures, called *EventSP* and *MarkupSP* respectively.

3.1 Principle of Visual Communication

It is always said that 'one picture is worth a thousand words' because we think the thousand words when we look at the picture. But it is not the right way to describe the visual contents in thousand words in our screenplay and it is not necessary. In fact images quickly fade from memory while the day's events are still within mind, so that a form of concept expression is suitable for us to write down events and ideas retained in our memory. An important issue involved is the possibility of automatically generated visual effects made of various media (e.g. animation, video) and modalities (e.g. music, talk).

Human beings perceive the world via the five senses of touch, hearing, sight, smell and taste. Film creates a five-dimensional world in the two-dimensional screen of sight and sound modes composed of different modalities. A modality indicates a particular form of a communication mode. For example, noise, music, and speech are modalities of the sound mode. For those modalities of smell or taste, their expressions in sound motion picture may be realized by speaking (in "rotted apple") or image (of rotted apple). Since the most important function of movie is to rightly express user's feelings, meanings and emotion toward audience, photorealism (realistic style in two respects: realistic picture or moving in realistic fashion) is not required.

3.2 EventSP and MarkupSP

EventSP (Event Screen**P**lay**)**

One kind of screenplay we designed is called EventSP. *Event* is an important primitive action unit in camera planning procedure such as "a person gives a speech" or "a private conversation between two characters" (See the example in Tab. 1).

Terms	Example	Note
Time	Daytime	When
Place	Sea park	Where
Character	A boy, a girl	Who
Prop	Trees	Which
Event	Two talk	What happened

Table 1. Two-talk Event Described in EventSP

There were some typical works on applying film theory for computer graphics generation. Christianson et al. adopted the notion of *film idioms* from film theory and formalized them into a sequence of shots [12]. He et al. encoded the film idioms into hierarchically organized finite state machine applied in real-time system [13]. Amerson & Kime proposed a system *FILM* (Film Idiom Language and Model) for real-time camera control in interactive narratives [14]. In our system, intelligent rule-based reasoning is employed which will be demonstrated in the next section.

MarkupSP (**Markup S**creen**P**lay)

The mind's picture is a combine of the perceptual elements of color, form, depth and movement combined with the verbal thoughts. To describe their imagery concretely, user should be allowed to add their controls in screenplay such as actions of characters (e.g. *stand*) or layout (e.g. on the left) in various shots (e.g. *close up*). These controls are included in filmmaking techniques involving the four aspects:

- *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 and
- *sound edition* (how to present the sounds), e.g., dialog, music, background sound from film theory.

4. E-moviemaking from Screenplay

4.1 EMM System Structure

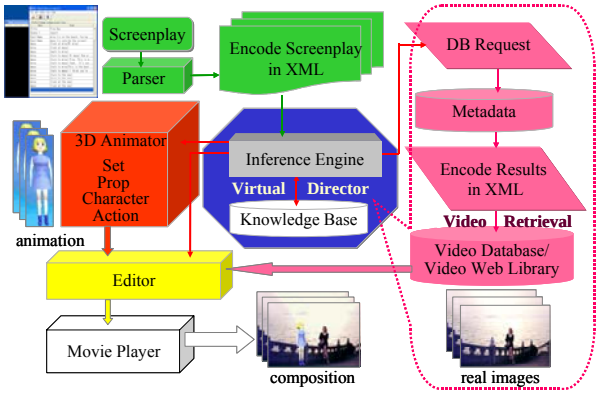


Figure 1. System Architecture Diagram

Virtual Director in Fig. 1 is embedded as a subsystem in the integrated system environment to realize the automation of 3D animation and video retrieval. He is responsible for the visual aspect of screenplay dependent on knowledge of plot structure in KB, giving commands for the dramatic structure, pace, and directional flow elements of the sounds and visual images to visualize the event. Supposing the existence of a library that stores 3D models and actions mentioned in the script, it is possible to combine objects and actions according to the screenplay and to choose optimal placement for the camera automatically. Composition, the location of characters, lighting styles, depth of field and camera angle are all determinant factors in the formulation of the visual information. Movie player assembles the resultant plan created by inference engine into images. *Virtual camera* records the frames that are to be played as a still or a sequence of images.

4.2 EMMVAR

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 2. Multi-modal Query in EMMVR

EMM (EMM Video Retrieval), a subsystem of EMM, focuses on design multi-modal video indexing. Giving an overview of EMMVR in one sentence, a suitable multi-category video modeling and multi-modal query (Tab. 2) mechanism with semantic video indexing using visual features, non-visual features, and sound cues were constructed based on MPEG-7 as well as from the point of view of film director.

Automated indexing approach is required if possible because fully manual video content indexing is a very time-consuming procedure. But *fully automatic semantic annotation* is still impossible with current VR technology. For the content that cannot be annotated automatically, *computer aided content indexing* may be chosen as a feasible way for complement. Their detailed explanations are showed in table 3.

Approaches	Tasks
Automatic annotation (Fig 2)	1. Segment (vs. montage): Scene → Shot → Keyframe. 2. Semantic feature extraction (vs. mise-en-scène): –Set, character, and prop in specific domain; –Some camerawork like pan; –Sound: music, dialogue, etc. 3. Event extraction (vs. mise-en-scène & sound edition): e.g., sport type.
Computer aided annotation (Fig 3)	User provides indices through interface of the software detector.

Table 3. Approaches of Video Content Indexing

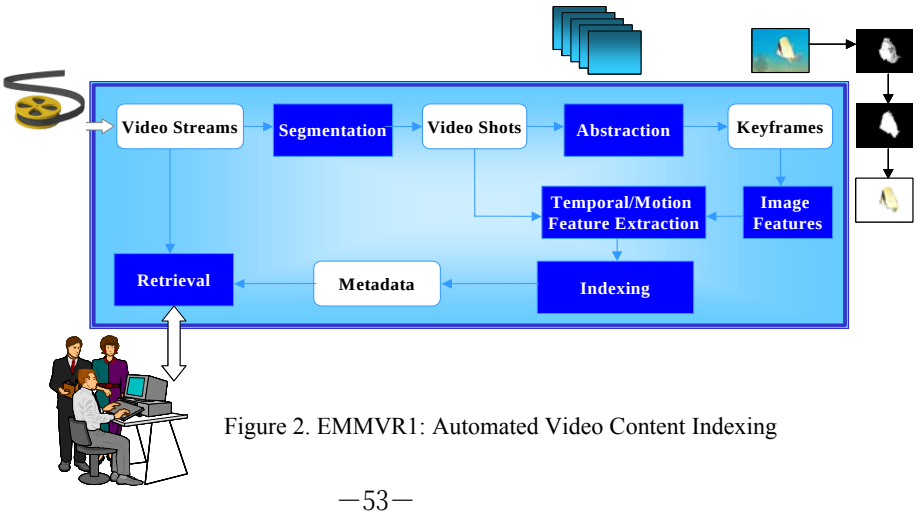


Figure 2. EMMVR1: Automated Video Content Indexing

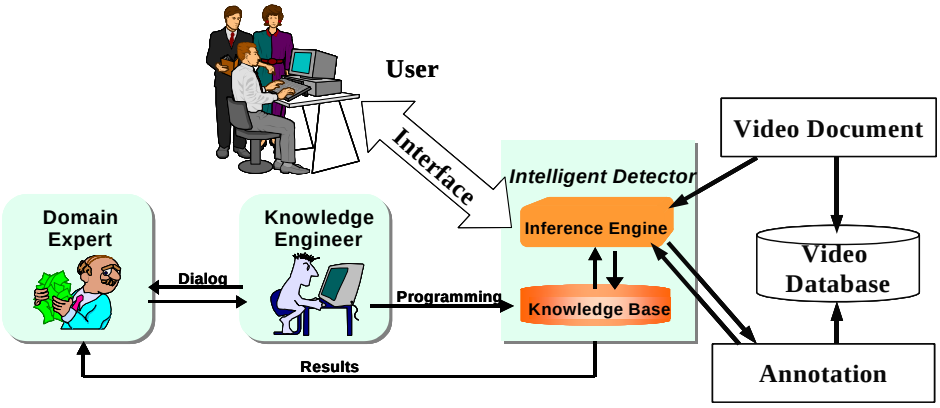


Figure 3. EMMVR2: Computer Aided Annotation

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. XML tags related to video contents are supported by those dark squares in Fig. 4, indicating the main contents that should be extracted from video in order to reuse the video.

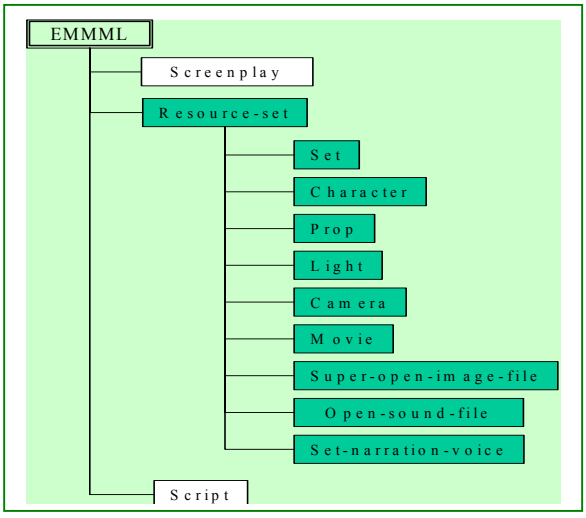


Figure 4. EMM XML Ontology Tree

4.3 How to Shoot

Categories of shots	Elements of Movement [ ] = may have, may have not
simple	*subject* (character, prop)
complex	*camera angular* (panning, tilt, rolling), *lens*, [subject]
developing	*Camera position*, i.e., mounting (tracking, dolly, crane), [lens], [camera angular], [subject]

Table 4. Categories of Camera Shots

Cinematography comprises camera angles, mobile framing and camera movements. Various definitions of shot are based on camera manipulation. We defined shot as the single uninterrupted operation of the camera that results in a continuous action. *Shot* such as *full shot, pan,* or *track* is the smallest unit of dramatic action in the movie. All of shots are grouped into three categories derived from the four elements of movement – subject, camera angular, lens, and camera position (Table 4). A complex shot and developing shot are showed in Fig. 5 and Fig. 6 respectively.

Figure 5. Lens (zoom in) + Pan + Tilt

Figure 6. Track Shot – Tracking One's Walk

A shot *sequence* is a group of shots depicting one action or which seems to belong with or depend upon each other. It is generated step by step to expound how to use cinematic 'rules of thumb' to make a scene. An example shot sequence is showed in Fig. 7 composed of shots transformed from medium shot to two-shot through dolly shot (concerning camera movment).

Figure 7. A Shot Sequence.

We model the filmmaking knowledge and rule-based reasoning strategies in expert system language CLIPS and embedding CLIPS into VC++.

4.4 How to Visualize Event

Heuristics about making a sequence of shots involves the techniques of montage and sound related to image, and another unit in film named event. For example, given dialog sentences of Tab. 5 in the event of two-talk introduced in section 3.2, the virtual director will stage face to face two-talk by the following steps: first, determines five basic shots from nine camera positions, then selects shots from the set and arranges them in order dependent on dialogues according to the planning rules (a) – (f) (involving continuity cutting). The reasoning procedure is showed is figure 8.

(a) If character A and B have a private conversation, five basic shots could be used: *two-shot* (default size: *full shot*), *profile shot* (default size: *close-up)*, *over-the-shoulder shot*, *point-of-view shot* (default size: *close-up)*, and *angular shot* (default size: *close-up)*.
(b) If both character A and B are silent, use two-shot.
(c) If character A talks, select one least used shot by A from the set of basic shots.
(d) If character B talks, select one least used shot by B from the set of basic shots.
(e) If character talks, OTS should be selected first.
(f) If the selected shot is not OTS, it should be set before OTS in the shot sequence.

Premises	Contents
(1) Silence	A: girl, B: boy
(2) B talks	Why did not you wear that yellow shirt that your sister gave you for your birthday.
(3) B talks	It looks terrific on you.
(4) A talks	I love the shirt, but it missed two buttons

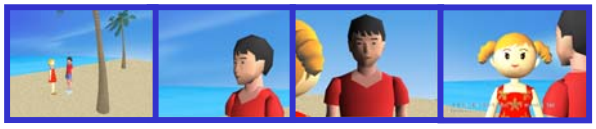
Table 5. Dialog in Two-talk Event

(1) —(b)⇒ **two-talk shot**

(2) —(e) (d)⇒ two-talk shot, **OTS shot (B)**

(3) —(f) (d)⇒ two-talk shot, { *profile shot,* / *point-of-view,* shot OTS shot (B) / **angular shot,** }

(4) —(e) (c)⇒ two-talk shot, angular shot, OTS shot (B), **OTS shot (A)**



1.Two-shot VLS 2.Profile-shot MS 3.OTS (facing A) 4.OTS (facing B)

Figure 8. Inference Procedure of Staging Dialogue

One thing must be noticed is that when the boy talks again, the virtual director stochastically selects a shot from the three shots of profile, point of view, and angular. That is to say there are other two possible results if profile was not selected. In the case when the size of shot is changed, there will be other new shot sequences like figure 9.

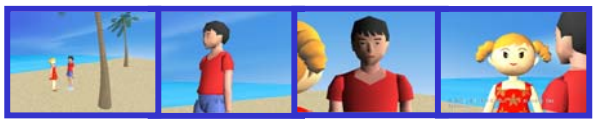


Figure 9. A Two-talk Shot Sequence

5. Discussion

In recent years, filmic techniques have been extended to a degree possible with live actors shot in real time, but the commercial software tools used to make CG movie are not common user access. For EMM, if there are suitable video clips in video database or video web library, the required clips will be extracted from the database/library, otherwise, 3D animation will be created based on cinematic knowledge, so that at present it is feasible to automatically make motion picture with visual effects like computer animation and real images, and their simple composition. The filmmaking knowledge base of our digital moviemaking system EMM contains domain knowledge about objects, color, lighting, scene, shot, also contains spatial-temporal knowledge. The cinematic knowledge-based environment instead of programming enables nonprofessionals to make their own digital movies easily.

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] Konstantinos Manos, Themis Panayiotopoulos, George Katsionis, "Virtual Director: Visualization of Simple Scenarios". 2nd Hellenic Conference on Artificial Intelligence, SETN-02, Thessaloniki, Greece, April 11-12, 2002

[4] Doron Friedman, "Yishai Feldman, Knowledge-Based Formalization of Cinematic Expression and its Application to Animation". Proc. Eurographics 2002, Saarbrucken, Germany, Sept. 2002, p.p.163-168,

[5] Kevin Kennedy, Robert. E. Mercer,"Planning animation cinematography and shot structure to communicate theme and mood". Proceedings of the 2nd international symposium on Smart graphics, June 2002, p.p.1-8.

[6] Szarowicz, A., Amiguet-Vercher, J., Forte, P., Briggs, J., Gelepithis, P.A.M., Remagnino, P., "The Application of AI to Automatically Generated Animation, Australian Joint Conference on Artificial Intelligence", AI'01, AI 2001:Advances in Artificial Intelligence, Adelaide, Dec 10-14, 2001, p.p. 487-494

[7] Stefan Kopp, Ipke Wachsmuth, "A Knowledge-based Approach for Lifelike Gesture Animation". In W. Horn, editor, ECAI 2000 Proceedings of the 14th European Conference on Artificial Intelligence, IOS Press, Amsterdam, 2000, p.p. 120-133

[8] John Funge, Xiaoyuan Tu, Demetri Terzopoulos, Cognitive Modeling: Knowledge, Reasoning and Planning for Intelligent Characters. *Computer Graphics Proceedings*, *Siggraph,* 1999, p.p. 29-38

[9] 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.

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

[11] Bob Coyne, Richard Sproat, "WordsEye: an automatic text-to-scene conversion system, Proceedings of the 28th annual conference on Computer graphics and interactive techniques", Aug. 2001, p.p. 487-496

[12] Christianson, Anderson, Wei-he, Salesin, Weld, and Cohen, "Declarative Camera Control for Automatic Cinematography". AAAI/IAAI, Vol. 1, Portland, Oregon, 1996, p.p. 148-155

[13] Li-wei He, Michael F. Cohen, David H. Salesin, "The virtual cinematographer: a paradigm for automatic real-time camera control and directing". Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, New Orleans, Louisiana, United States, August 1996, p.p. 217-224

[14] Amerson, D. and Kime, S., "Real Time Cinematic Camera Control for Interactive Narratives". In the Working Notes of the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment, Stanford, CA, 2001