A Knowledge-based Moviemaking Approach

JINHONG SHEN[1] SEIYA MIYAZAK I[2] TERUMASA AOKI[1] HIRHOSHI YASUDA[1]

[1] The University of Tokyo, 153-8904 Tokyo, Japan[1]

[2] Matsushita Electric Industrial Co., Ltd., 140-8632 Tokyo, Japan

Abstract: We are developing an intelligent approach of motion picture generation for desktop software system ***EMM*** (Electronic MovieMaker) that aims at automating the production of digital movies with various visual effects like three-dimension animation, real image, and their composition. This paper describes our efforts on EMM focusing on automatic synthesis of character animation and automatic video retrieval from video database/Web video library dependent on filmmaking rules, both of which have just right inverse processes opposed to each other. Two kinds of screenplay' user interface are introduced along with their theoretical basis of design. Then expounds how to visualize the screenplay by using cinematic 'rules of thumb' to make a scene. The implementation of building knowledge representation includes scene's layout and shooting and character's action written in *CLIPS* language.

Keywords: Electronic Movie Making, Cinematic Knowledge-based System, 3D Animation, Virtual Director, Content-based Video Retrieval

1. Introduction

Current computer technology has reached a level that allows us to create a virtual world we can imagine, but there is still a great difficulty in imagery creation for the reason that the process of generating dynamic 3D Computer Graphics is quite troublesome and time-consuming even if employing specific hardware and software, and requires user the high-trained skill and specific talent so that it is still impossible to produce personal movie by computer readily within short time when we come up with an idea for movie due to a variety of limits existing in these approaches: the time, the labor, the material, etc, these make the moviemaking on computer very expensive.

To cover the need of simplifying the process of digital movie making, one reasonable proposal is to develop an easy-to-learn and easy-to-use desktop software tool by which a general person can make his own visual contents and deliver it easily over phone lines or fiber-optic superhighway. After analyzing the feasibility of realization, we came to the conclusion that it is reasonable to automatically visualize a verbal screenplay by using relevant sound motion pictures with visual effects like real images, 3D animation, or their composition, where real images are extracted from digital video (movie, animation, TV programs, etc.) library [1], [2].

The software system EMM (Electronic Moviemaker) we are implementing for such automated digital moviemaking may generate shot sequence 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*). 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 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 Japanese NHK's TVML Language to show on TVML Player 1.2.

The next section describes related works on automated language-based movie generation system and video retrieval, and a statement of the issues. Section 3 first analyzes the functionalities of computer animation and video retrieval, then puts forward new filmmaking techniques from film theory utilized in the virtual 3D world and gives a conceptual framework of knowledge-based approach for realizing automatic digital moviemaking. In section 4, we explain how to design knowledge representation including scene's layout and shooting and character's action written in *CLIPS* language. Two kinds of screenplay' user interface are introduced along with their theoretic basis of design. Later show the system implementation with pieces of animation to expound how to use cinematic 'rules of thumb' to make a scene. Finally, I will summarize the contributions of our present work and discuss about future work.

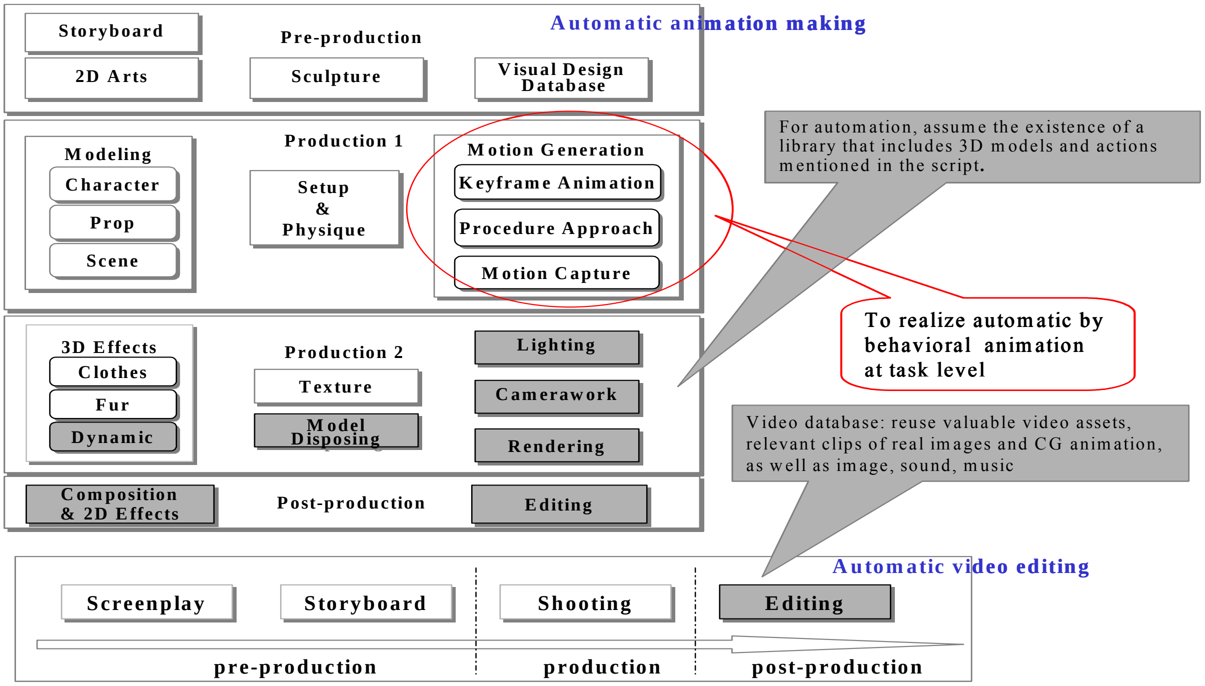
2. Related Researches

The production system EMM 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. Those 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 moviemaking 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].

EMM also uses content-based retrieval techniques to automate the procedure of video retrieval. 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 systems (Photobook, VisualSEEK) and commercial systems (QBIC, Virage) 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].

3. Feasibility Analisis of Automation

Both of the production phases of CG (Computer Graphics) movie/computer animation and traditional film/video are divided into pre-production, production, and post-production as grouped in figure 1. For computer animation, 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. For digital video, the production using DV camera encompasses acquisition, storage, selection/editing and composition of video data. Except that actual shooting requires human's



For automation, assume the existence of a library that includes 3D models and actions mentioned in the script.

To realize automatic by behavioral animation at task level

Video database: reuse valuable video assets, relevant clips of real images and CG animation, as well as image, sound, music

Fig 1. Automation in EMM

involvement, the process of video choosing and sequencing can be automated based on experienced editing knowledge. That is to say the process of digital video production is at most of automatic edition and composition. Therefore automatic edition and computer animation are feasible. The dull rectangles in figure 1 indicate the functions that can be realized automatically.

Table 1. 3D Computer Animation Techniques

Stages	Related Technologies	Use AI?
1.Modeling	*Surface modeling **Polygonal surface **Curved and patched surface *Solid modeling *Particle system modeling for fire, cloud, mist, spray smoke, and so on	
2.Animation/ Motion generation	*Modeling-based **Geometric (Kinematical)/ Keyframe-based -Hierarchical animation --Forward kinematics --Inverse kinematics -Shape deformation **Procedural/rule-based -Particle system -Physically-based simulation --Passive system --Active system **Behavioral ------------------- *Motion capture-based **Magnetics-based **Optical-based	Yes
3.Rendering	*Volume rendering *2D image manipulation, and so on	

Professionals conduct 3D computer animation production through modeling, animation and rendering. Stages of modeling and animation require user's detailed expert knowledge on animation making and are quite time-consuming. 3D computer animation or motion generation techniques include *kinematical* (*geometric*), *procedural* (rule-based), *behavioral*, and *motion capture* approaches (table 1). The former three approaches are classified to *model-based* approaches among of which *behavioral approach* heavily relies on the techniques of AI and is built on other motion generation techniques such as *physically-based simulation* and *inverse kinematics*, so that it is a feasible way to realize automatic animation by behavioral approach at task level. Table 1 shows the AI sub-field to which behavioral approach belongs and others main applications of related AI technologies.

4. E-Moviemaking

4.1 System Structure

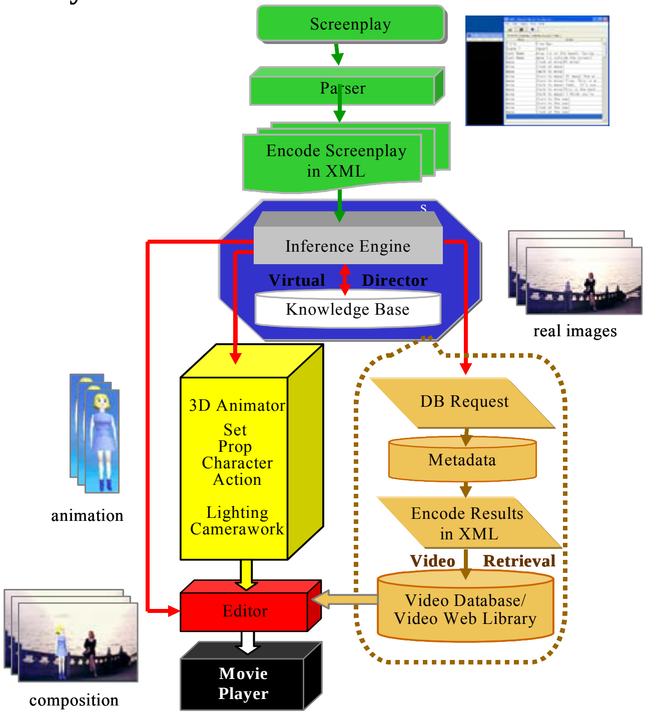


Fig 2. *EMM System Architecture Diagram*

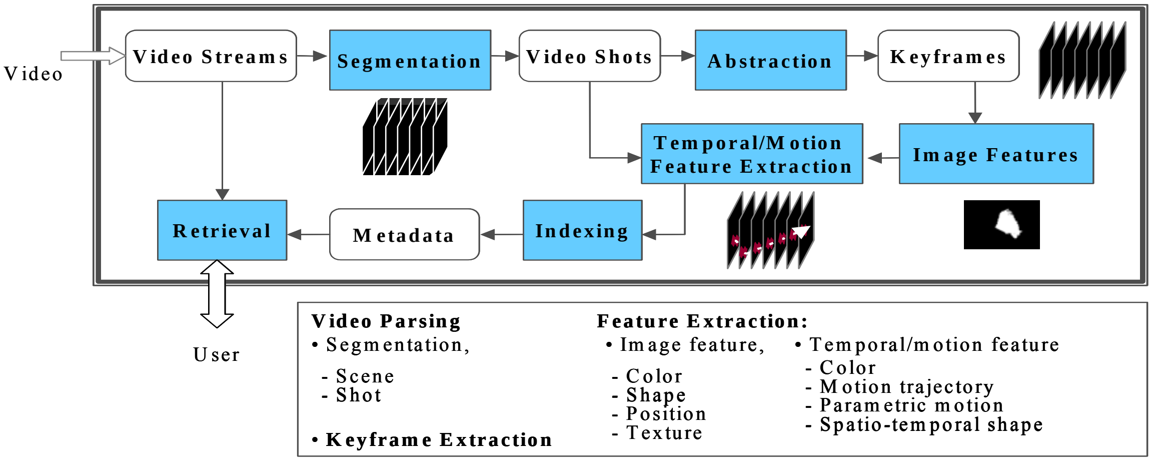


Fig 3. EMMVR1: Automatic Annotation

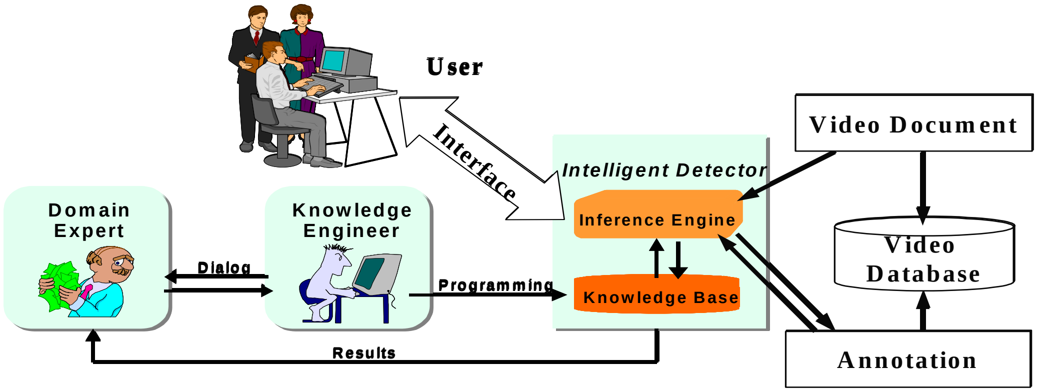


Fig 4. EMMVR2: Computer Aided Annotation

The whole EMM system structure showed as figure 2 is an integrated system environment. To realize the automation of 3D animation and video retrieval, the system core *Virtual Director* is responsible for the visual aspect of screenplay dependent on knowledge of plot structure in KB. He gives 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. *EMMVR (EMM Virtual Retrieval)* has a suitable multi-category video modeling which can represent necessary semantic, syntactic and structural information as well as from the film director's perspective and multi-modal query mechanism which supports querying by example and text like keyframe). Its design focuses on multi-modal video indexing constructed based on MPEG-7. Automated indexing approach is required because fully manual video content indexing is a very time-consuming procedure. But fully automatic semantic annotation is still impossible by current VR technology. For the content that cannot be annotated automatically, *computer aided content indexing* may be chosen for complement (table 2).

Table 2. Approaches of Video Content Indexing

Approaches	Tasks
Automatic annotation (Fig. 3)	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. 4)	User provides indices through interface of the software detector.

4.2 Screenplay

Among the works on translating verbal presentations into visualized presentations, AT & T is making a system named WordsEye [11] for automatically converting text into representative 3D static scenes. However, though natural language is an easy and effective medium for describing visual ideas and mental imaginary, 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.

In our system intelligent filmmaking rule-based reasoning is employed. There are 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]. Dependent on cinematic knowledge base the digital filmmaking procedure can be further automated by Artificial Intelligent approach. For those nonprogrammers and non-artist, the cinematic knowledge-based environment instead of programming will save them time and labor greatly.

When we design 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 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. We chose screenplay as input because it is a formal language for filmmaking, which implies the lots of rules of film that are almost invisible by audience.

4.2.1 EventSP (Event ScreenPlay)

When we carefully analyze a visual message, we find images become real property of the mind and are remembered only when language expresses them. Symbols are used to retain events and ideas in our memory so that one kind of screenplay we designed is called EventSP that can describe abstract relationship between objects. An event description should at least contain information like time/when, place/where, people/who, prop/which as well. These information need to be input by user such as the following *two talk* event input information.

Event two-talk
Time day
Place park
Prop plam
Character boy Jack
Character girl Marry
Talk Jack "Why don't you wear that yellow shirt that your sister gave for your birthday?"
Talk Jack "It looks terrific on you."
Talk Marry "I love the shirt. But it's missing two buttons."

4.2.2 MarkupSP (Markup ScreenPlay)

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.

When designing the screenplay format, an important issue that must be considered 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 modalities of smell and taste, their expressions in sound motion picture may be realized by speaking ("rotted apple") or image (rotted apple). Since the most important function of movie is to rightly express user's feelings, meanings and emotion toward audience, photo-realism (realistic style in two respects: realistic picture or moving in realistic fashion) is not required.

4.3 How to Shoot

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* (showed in Fig 5) is the smallest unit of dramatic action in the movie. The virtual camera is modeled with seven Degrees of Freedom (DOFs) - three for Cartesian position, three for orientation, and FOV (Field of View) all the same as those a common real-world camera has, so that it could be controlled in the same way as the real one.Dynamic picture generation concerns

1) object description,
2) object movement, and
3) camerawork-shot and shot sequence.

CLIPS provides a cohesive tool for handling a wide variety of knowledge with supports for three different programming paradigms: rule-based, object-oriented and procedural so that it can fulfill the above mentioned needs for high-level tool to program the generation of digital movie.

Event is an important primitive action unit in camera planning procedure such as "a private conversation between two characters". To stage the example event of face to face two-talk (the boy Character B talks two sentences, the girl Character A gives one sentence answer), the virtual director first determines five basic shots from nine camera positions, then selects shots from the set and arranges them (Table 3) in order dependent on dialogues according to the following planning rules.

(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)*. (Fig 6)

(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.

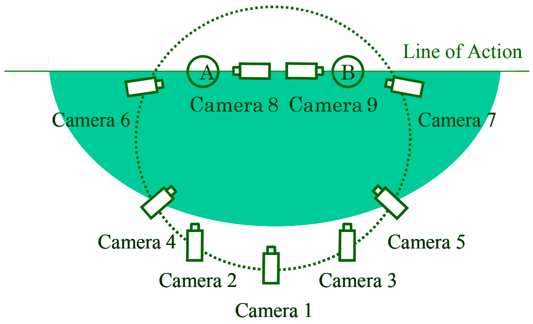
(e) If character talks,
OTS should be selected first.

(f) If the selected shot is not OTS,
it should be set before OTS in the shot sequence.

(g) If it is the first shot,
layout two trees on the right side, one on the topand one on the bottom,

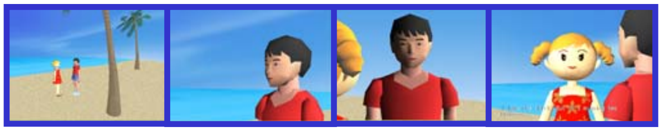
layout character A on the left and character B on the right, set bright light,

both characters are silent.



Camera 1 : two-shot (default size: full shot),
Camera 2, 3: profile shot (default size: close-up),
Camera 4, 5: angular shot (default size: close-up),
Camera 6, 7: over-the-shoulder shot,
Camera 8, 9: point-of-view shot (default size: close-up)

Fig 6. Camera Placements in Two-talk Sequence



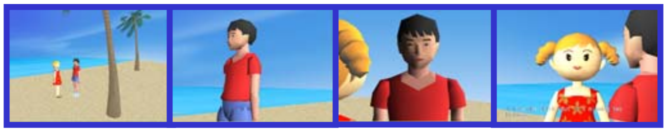
1.Two-shot VLS 2. Angular-shot CU 3.OTS (facing A) 4.OTS (facing B)

Fig 5. *Two-talk Shot Sequence*

Table 3. Staging Dialogue Sequence for Two Characters

Inference procedure		
	Actions	
Premises	By rules	Result
Two talk	(a)(g)	Set of basic shots of two-shot & very large shot (VLS) profile-shot & close-up (CU) over-the-shoulder shot (OTS) point-of-view shot (POV) & close-up angular shot & close-up
Silence	(b)	Sequence of shots **1. Two-shot VLS**
B talks	(d) (e)	Sequence of shots 1. Two-shot VLS **2. OTS (facing B)**
B talks	(d) (f)	Sequence of shots 1. Two-shot VLS **2. Angular-shot CU** 3. OTS (facing B)
A talks	(c) (e)	Sequence of shots 1. Two-shot VLS 2. Angular-shot CU 3. OTS (facing B) **4. OTS (facing A)**

Besides the sequence of Fig 5, there are other possible results according to the above planning rules (a) – (f) because of some stochastic process in shot selection (e.g., shot sequence of 1.Two-shot VLS, 2.Profile-shot CU, 3.OTS (facing A), 4.OTS (facing B)). If adding new shot sizes to the shots maybe used, there will be more possible shot sequences resulted for the same event such as the result in figure 7.



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

Fig 7. Two-talk Shot Sequence

In rule-based language, a rule is a concise description of a set of conditions and a set of actions to take if the conditions are true. Film rules about above shooting can be written in *defrule* construct of CLIPS as rule (a):

```
(defrule TwoTalk “A rule of intention shot”
  (Event two-talk)                                  ; If
  =>                                                ; Then
  (assert (Shot two-shot VLS))
  (assert (Shot profile-shot))
  (assert (Shot OTS CU))
  (assert (Shot POV CU))
  (assert (Shot angular-shot CU)))
```

where two parts are separated by the ‘=>’ symbol (means ‘then’). The first part consists of the LHS left-hand side *pattern* (track-two-half-front) which is used to match facts in the knowledge base while the second part consists of the RHS right-hand side *actions* that contain function calls. The rule of shot selecting will be activated when the fact (Event two-talk) appear in the knowledge base. When the rule executes or *fires*, the functions (assert (Shot two-shot VLS)), (assert (Shot profile-shot)), (assert (Shot OTS CU)), (assert (Shot POV CU)), (assert (Shot angular-shot CU) are called. Annotation begins with symbol ';'.

We have demonstrated to encode spatial constraints of a shot such as over-the-shoulder in CLISP that concerns the domain knowledge about the techniques of mise-en-scène and cinematography. CLIPS codes of shot and rules presentation about shot sequence may be referred in the paper [1].

5. Discussion

A low-cost easy-to-use electronic moviemaker has good entertainment and education marketplace. Making such a system is a great challenge. The filmmaking knowledge base of our digital moviemaking system EMM contains domain knowledge about objects, color, lighting, scene, shot, also contains spatial-temporal knowledge. 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. EMM employs numerous cameras that the virtual director or user controls. The cameras that follow director/user's control can be set more than one and at different positions and varying speeds. EMM has plentiful and interesting results. For the same event, virtual director can "image" many scenes. Instead of data translation, EMM supports semantic translation. Knowledge representation subsystem and other database subsystems can be embedded and over which ontology is defined. The cinematic knowledge-based environment instead of programming enables nonprofessionals to make their own digital movies easily.

Reference

[1] Jinhong SHEN, Seiya MIYAZAKI, Terumasa AOKI, Hiroshi YASUDA, Intelligent Computer Moviemaker with the Capabilities of Cinematic Rule-based Reasoning (II), The Journal of the Institute of Image Information and Television Engineers (ITE), (Tokyo, Japan, July 2004), Vol. 7, 974-981

[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), (Florida, USA, Aug 2003), VI, 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), 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), 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), 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, (Amsterdam, 2000), IOS Press, 120-133

[8] John Funge, Xiaoyuan Tu, Demetri Terzopoulos, Cognitive Modeling: Knowledge, Reasoning and Planning for Intelligent Characters. Computer Graphics Proceedings, **(**Siggraph, 1999), 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), 487-496

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

[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), 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)