オーディオビジュアル複合情報処理　39－21
(2002. 12. 17)

The Framework of an Automatic Digital Movie Producer

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 The aim of our project is to develop an easy-to-use tool that provides an integrated environment for dealing with the complete process of creating digital film. To make the creating process automated, we propose to employ a verbal screenplay as input form. The system we designed can understand the script through a parser, then a virtual director and a virtual cinematographer automatically translate it into a relevant motion picture with various visual effects like real image, three-dimension (3D) animation, or augmented reality. In this paper, "video" stands for all kinds of clips of these visual presentations. On account of the ability to extract suitable video clips from digital video web library and present movie production on web, video data in our system is encoded in XML and tracked by the MPEG-7 standard. This system is an extension of DMP [Seiya and others, 2002].

Keyword 3D animation, virtual director, metadata, knowledge base, XML

1. Motivation

Though current computer technology has reach a level that allows professionals to create a virtual world they can imagine, there is a great difficulty with interactive point-click 3D animators and filmmakers because the creating process of generating 3D animation is quite troublesome and time-consuming, as well film editing. When using animators (e.g., Alias ***Maya***, Autodesk ***3D StudioMax***, Avid ***SoftImage***), users need to own knowledge of mathematics, computer 3D techniques, and artists to understand a complex software package. They keep on traversing control widgets and transforming parameters continuously. In addition, they must undertake a lengthy off-line programming session. Video editors (e.g., Adobe ***Premiere***, Apple ***iMove***) also remain difficult because digital video is time-based medium having dual tracks of audio and video. We are exploring to implement an easy-to-learn and easy-to-use tool for making digital movies contain various visual effects such as real images, 3D animation, or augmented reality. Our research precisely focuses on simplifying the process of movie making.

The rest of this paper is organized as follows. In section 2, we mentions related works in automated language-based movie generation system. After that, we describe our approach for automatically making video clips of real effects, animation, or augmented reality in an integrated environment from a formal script. We conclude with a show of the contributions of the system and future works.

2. Related Works

2.1. Verb-to-image

Emerging issue of designing such a system is on human-computer interface. We think current desktop window-oriented GUI is not always well suited to computer vision since human-computer interaction and usability often reflect the engineering technologies rather than simply for users. Another disadvantage is that more abstract input commands such as conditions are short in interactive control. In our design, computer interface must evolve to let us utilize more of the power of language.

Some works on translating verbal presentations into visualized presentations are in progress. At & T' is making a system named WordsEye [Bob and others, 2001] for automatically converting text into representative 3D 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. On the other hand, 3D animation is far more difficult to be realized than 3D scene, depending on synthetic techniques involving the fields of Linguistics, Artificial Intelligence, Computation, and Computer Vision.

We propose to employ verbal screenplay as input form so that it is not necessary to understand full natural language. We know the language of film is a formal one that has evolved gradually through the efforts of talented

filmmakers from the beginning of the century. It implies the lots of rules of film that are almost invisible by audience.

2.2. Video formats

Digital video production encompasses the acquisition, storage, selection/editing, and composition of video data. Most efforts in this area have focused on development of video database systems low-level tasks such as video segmenting, indexing, modeling, querying, and editing. A video database may contain films, news broadcasts, archive footage, etc.

Video can be considered as structure documents with various granules from sequence or story to low-level audiovisual objects (e.g., face). Effective segmentation of images into visual objects is very helpful for the further video segmentation and interpretation of the resulting granules. Recent standardization efforts such as MPEG-4 are based on the capability to automatically segment video frames into objects. They rely on mainly on movement-based segmentation so that is not always available and effective. Video processing systems compatible with MPEG-7 can realize the representation of audiovisual contents via various level of semantic objects, and the description of the content of these objects.

2.3. Automatic editing

Except that actual shooting requires human involvement, the process of video choosing and sequencing can be automated. There are several systems (e.g., VRSS, Hitchock) that can edit video with varying degrees of automation but no such a system that has professional result. Fully automated systems only may be used for news on demand.

2.4. Automatic animation

Relatively little effort has focused on the task that automatically converting text into representative 3D animation. Efforts on automatic animation are focusing on automatic synthesis of character animation. The goal has many difficult problems such as how to generate primitive motion for any class of action (run, jump, etc.), how to parameterize these actions at a high level, and how to combine the primitive motion to create coherent sequences of actions. Most approaches to automatic motion synthesis generally appeal to some combination of physics simulation and robotic control to generate motion from a high level description. Advanced algorithm is based on motion transformation, which produces new motions by combining/deforming exiting motions.

2.5. Augmented reality

Augmented reality is the technique by which real images can be enhanced by addition of computer-generated or real information. This area up-to-now been largely under-explored and the few applications have been developed rely heavily on the intervention of the user at the different steps of the mixing process.

3. Design of an automatic movie maker

3.1. The data format

(1) Script form In our system, the script is not arbitrary general-purpose natural language text, but looks similar to the verbal screenplay for film/telefilm so that it is more accessible to non-programmer user. See an expert of film

Location: Research room 555.
Jack stands in front of whiteboard.
Tom entered and speaks: "Jack, have you eaten the apple I pared?"
Tom sits on sofa.
Jack speaks: "I was hungry."

We can see that screenplay typically describe specific characters performing concrete actions rather than only abstract relationship.

XML (extensible Markup Language) is employed to format information of script since XML acts as a description language for multimedia well. For example,

<sequence name="Shot1" firstimage="007" lastimage="101" .../>

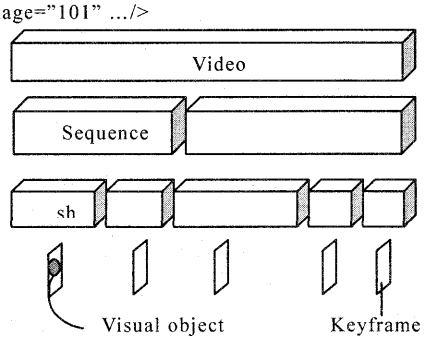

Figure 1. A hierarchy structure of video

(2) Hierarchy structure of video Our video has a hierarchy structure (Fig. 1). At the highest level, a video is a sequence of scenes, each of which captures a specific situation or action. Each scene is composed of one or more shots. A single shot is the interval during which the video camera is rolling continuously. Still image is called frame and a keyframe is such a frame that represents

video sequence. In standard approaches, the analysis of the various temporal granules of the hierarchy is done in a bottom-up manner while much high-level semantic knowledge can be inferred by a top-down analysis.

(3) Video database The system obtains metadata from the ***Informedia*** Digital Video Library. This is a growing searchable multimedia library that currently has over 2000 hours of material, including documentaries and news broadcasts. **Informedia** adds two hours of additional news material every day. Video model of metadata is encoded in XML and tracked by the MPEG-7 because XML is a language suited to describe structured information and their properties. For example,

<element name="Talk" type="string"/>

3.2. Theoretical foundation

System does not target at a modeling package of characters, props, or scenery, as there are many systems that already do this. Assume the existence of a library that includes 3D models and actions mentioned in the script. The system needs to combine objects and actions according to the script, to convert text to speech and to draw from a library of sound effects and background music. Though interactive 3D computer graphics applications (e.g., virtual chat managers, fiction environment, video games, and virtual agent) could be set up on specialized training and animation skills by utilizing above mentioned animation tools in section one, there is a great limit to camera actions. The reason is that some idioms, usually portrayed from a particular character's viewpoint or from a small set of strategically-place points of view, are used in automata for virtual camera control so that the computer animation results in some fixed modes, hardly exploiting established cinematographic techniques.

(1) Ontology For human beings, information that we encounter is understood through our own internal data structures, which are our own implicit organizations of knowledge retrieved from the world we inhabit. With regards to information processing in machines, automatic movie generation system also needs an ontology that possesses sufficient semantics for making movie from script. Knowledge is without meaning unless it is contextualized. Ontology can be seen as a conceptual map where the links between individual pieces of knowledge are delineated. A precise manner is needed to encode the large body of cinematic knowledge into knowledge base for computer to manipulate.

Rules in making film are usually about the following three aspects. **(1) Camera control** closeup, closeup, medium view, full view, and long view. **(2) Heuristics for selecting shots** for example, avoid jump cuts. A cross the cut there should be a marked difference in the size, view, or number of actors between the two setups. A cut failing to meet these conditions creates a jerk, sloppy effect. **(3) Editing film** dissolve, fade in/out, etc. For the same clips, different sequences of shots lead different styles.

(2) Virtual director For automated filmmaking, a process would be analogous to creating a virtual director in place of a human one. Information about common film idioms may be encoded using a special narrative (e.g. TVML) [Hayashi and others, 1999]. The direction object takes the translated input and determines which scene is described in the narrative best fit the script. Once a scene is selected, the director binds any unbound variables may be time constants, locations, rotations, actors, or props in the world.

(3) Virtual cinematographer Once the director has specified the scene and the constraints that effect the filming of that scene, it is the cinematographer's job to choose the optimal position for the camera. The cinematographer uses the constraints to select an optimal placement for the camera, but this selected optimal may or may not be a valid position in the virtual environment. Some possible reasons that can make an optimal position invalid are the shot may be occluded by another object in the world or a pair of the constraints may not be easily satisfied in tandem. In this case, we relax the constraints out from our optimal point determine the closest point that will best satisfy our constraints. Constrains are relaxed based on the weights defined in the scene specification.

3.3. Prototype of system

The guiding principle of all aspect of our work is that we can make automatic motion synthesis tractable by attaching expert knowledge to otherwise unstructured example motion. This knowledge takes the form of descriptive annotations on the example motions and rules for how to combine and transform example motions. We outline the prototype of our system in Fig 2.

The system allows of processing a various types of scripts of movies. Then produces relevant digital films on the basis of cinematic principles. If there are suitable video clips in video database, the required clips will be extract from the video database or video web library. Otherwise, 3D animation will be made by virtual director and cinematographer based on expert knowledge base. If

necessary, augmented reality composition will be produced.

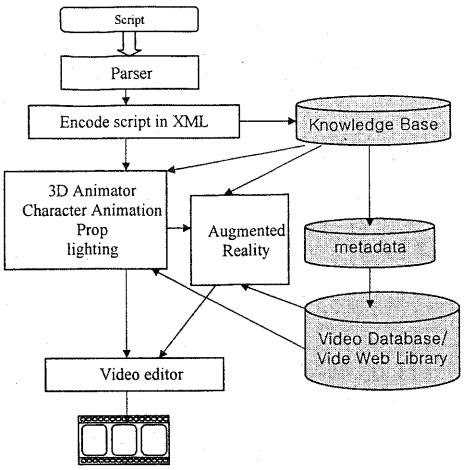

Figure 2 Structure of an Automatic Digital Movie Producer

For automated character animation, the approach consists of three parts.

(1) A toolkit and methodology

It is used for motion transformation that combines the strengths of previous transformation operators and extends the range of deformations apply to any set of motion. Blending, space warping, and time warping are existing operators. We present updates to these existing operators.

(2) Motion model

It is an abstraction for encapsulating domain-specific rules relevant to a particular type of action. We can reason about and manipulate gross aspect of the motion since we use transformation and example motions containing the details of the motion.

(3) Combining Algorithm

It is an algorithm for combining the primitive motion produced by motion models.

4. Conclusion

Our research on DMP is still in its infancy, and there are lots of concrete issues and technologies in the relevant areas need to be addressed and mingled. This embodies the creative thought and full potential of the project in forward position.

We have many works need to do in the future. The system needs to decide which objects (animation or reality) will participate, their location and design (color, texture, state), and the exact timing of actions. This stage requires a relevant knowledge base and involves Augmented Reality. It is possible to extend the augmentation of a screenplay with camera directions to modeling lighting and soundtrack.

Reference

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

[2] B. Tomlinson, B. Blumberg, and D. Nain. "Expressive autonomous cinematography for interactive virtual environments." In Proceedings of the fourth international conference on Autonomous agents, pages 317--324. ACM Press, 2000

[3] Casares, J. Myers, B, Long, A. C., Bhatnagar, R. Stevens, S., Dabbish, L., Yocum, D. and Corbett, A. "Simplifying Video Editing Using Metadata," in Processings of Designing Interactive Systems (DIS 2002), London, UK, June 2002.

[4] C. Roisin, T. Tran Thuong, L. Villard. "Integration of structured video in a multimedia authoring system." Proc. of the Eurographics Multimedia'99 Workshop, Springer Computer Science, ed., pp. 133-142, Milan, Sept. 1999.

[5] David B. Christianson, Sean E. Anderson, Li-wei He, David H. Salesin, Daniel S. Weld, and Michael F. Cohen. "Declarative camera control for automatic cinematography." In Proceedings of the AAAI-96, Aug. 1996.

[6] Don Gentner & Jakob Nielsen. "The Anti-Mac Interface." Communications of the ACM, Vol. 39, No. 8, pp. 70-82, Aug. 1996

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

[8] Hayashi, M., Ueda, H., Kurihara, T, Yasumura M., "TVML (TV program Making Language) - Automatic TV Program Generation from Text-based Script -", Imagina99 proceedings, 1999.

[9] He, L., M. F. Cohen, D. H. Salesin, "The Virtual Cinematographer: A Paradigm for Automatic Real-time Camera Control and Directing," in Proceedings of SIGGRAPH 96, Computer Graphics Proceedings, Annual Conference Series, 217-224,. August 1996),

[10] M.J. Conway. *Alice: Easy-to-Learn 3D Scripting for Novices*. PhD thesis, School of Engineering and Applied Science, University of Virginia, December 1997.

[11] Seiya Miyazaki, Terumasa Aoki, Hiroshi Yasuda, "DMP - an innovative personal Digital Movie Producer system", JPSJ, 2002-CG-108, Aug.2002.

[12] Ronald Baecker, Alan J. Rosenthal, Naomi Friedlander, Eric Smith, Andrew Cohen "A multimedia system for authoring motion pictures," Proceedings of ACM Multimedia' 96, pp 31 – 42, Nov. 1996.