Supporting the Creation of Immersive CG Contents with Enhanced User Involvement

Masashi OKAMOTO[*]
Kazunori OKAMOTO[*]
[*]Graduate School of Information Science and Technology, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
okamoto@kc.t.u-tokyo.ac.jp, kazu@kc.t.u-tokyo.ac.jp

Yukiko I. NAKANO[†]
[†]Research Institute of Science and Technology for Society, Japan Science and Technology Agency
Atago Green Hills MORI Tower 18F, 2-5-1 Atago, Minato-ku, Tokyo, 105-6218, Japan
nakano@kc.t.u-tokyo.ac.jp

Toyoaki NISHIDA[**]
[**]Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
nishida@i.kyoto-u.ac.jp

Abstract

As computer technologies have been advanced recently, people tend to feel stressed against poor or unsatisfactory interaction with computers. In order to solve such a situation we propose the human-computer interaction design theory of 'User involvement', which is expected to give guidelines for constructing a natural interaction environment for humans and computers. Specifically, we show the effectiveness of our theory in building a support system for immersive CG contents.

1 Introduction

Nowadays as computer technologies have been advanced rapidly, we often encounter the scene where people feel stressed against poor or unsatisfactory interaction with computers. For instance, in operating a word processor or in communicating with interactive QA system, the users are frequently annoyed with a cumbersome agent or irrelevant responses from the system. On the other hand, interesting TV programs and computer games do not lose their popularity from their watchers and players. What makes the difference?

Our answer is that it is because many of computer-mediated interactive contents leave the cognitive involvement of their users out of consideration that such problems occur. We do not want to insist that the users should participate in designing computer programs. Instead, the designers and the engineers should make out the cognitive activity of the users who use computers feeling *reality* in the virtual world before their eyes or in human-to-computer interaction in which they are engaged. On these occasions the users are deeply into another world or willingly keep communicating with virtual agents or robots. In other words, they are *involved* in the human-to-computer interaction environment.

We thus call such a cognitive activity or state of computer users 'User involvement' (Okamoto et al., 2004). Considering the user involvement helps to make an affective and favourable design of human-to-computer interaction environment.

In this paper, we first propose the idea of user involvement in brief. Specifically, we place much emphasis on the relations of reality and empathy based on cognitive linguistics research. And then a cognitive model for movie contents is shown based on observations on TV programs in Section 2. Lastly, as an example of the systems with enhanced user involvement, our ongoing work on a supporting system for the immersive CG contents creation is described. It is built to support a user who wants to create CG contents with little effort.

2 User Involvement in Human-Computer Interaction

The connotations of 'human-computer interaction' range from an explicit communication with a conversational agent or an intelligent robot to implicit interactions in playing computer games or watching CG contents. When a user consciously communicates with a robot, it can be said that he is *engaged* in that communication in the sense of Sidner et al. (2003). Thus the *engagement* might be applied to explicit human-computer communication.

In this paper we use '*(User) involvement*' instead so that it should include the situation where a user is involuntarily involved in a human-computer interaction or a virtual world. Our research focuses on finding out how the user involvement can be established and enhanced through a good design of human-computer interaction environment.

2.1 Requirements of user involvement

Our definition of the 'User involvement' and the main requirements to establish it are as follows:

User involvement. The cognitive way humans willingly engage in, or are forced to be involved in a virtual world which computers display, in a human-to-robot communication, or in a computer-mediated community.

Requirements.

1. Cognitive/Communicative/Social reality should be achieved.
2. Two (or more) cognitive spaces should be linked, and the user should cognitively move in and out those spaces.

In this approach the 'reality' is classified into the following three dimensions:

- **Cognitive reality.** The way of seeing objects, events and their relations in the real/virtual world as real.
- **Communicative reality.** The sense of reality that is achieved through communication with others.
- **Social reality.** The collective and intersubjective sense of reality based on sharing thoughts or opinions with one another.

In this research we mainly focus on establishing cognitive reality and communicative reality because social reality is considered to concern computer-mediated communities or online communities on the Internet.

2.2 Astigmatic model of user involvement

Since the user involvement is strongly related to our sense of reality and is common to both verbal and nonverbal communications, linguistics researches help to comprehend how it works. In particular recent cognitive linguistics suggest many important characteristics of human cognition in conceptualizing the world.

Langacker (1993) points out the reference-point ability, which enables us to conceptualize an entity at a distance using a mental path from a more accessible entity as a reference point. Applying this concept to the user involvement, it can be said that people conceptualize an unfamiliar entity in another world utilizing a more accessible one as a reference point that stands on both our world and the other.

At the time both of the worlds (i.e. cognitive spaces) need to be linked or overlapped with the reference point. That is not limited to the relation between a real space and a virtual space. At the very start of our life, we are living in two spaces, that is, thinking in the inner space and acting in the outer world using our body as a reference point.

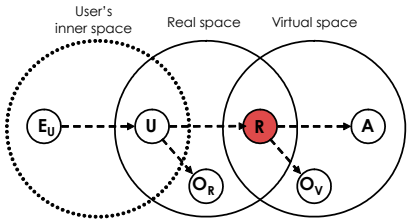

Figure 1: Astigmatic model

Figure 1 shows an example of how the computer user conceptualizes each object in different cognitive spaces using each reference point, especially in a virtual agent system. We call this model as the astigmatic model of user involvement, because multi-spaces are overlapped and linked there.

A computer user (U) accesses an object in a virtual space (O_v) via some reference point (R) that is easily accessible for the user and, at the same time, is a constituent of the virtual space. For instance the correspondent movements of a mouse and its pointer function as a reference point in that it connects our real world and a virtual world in the computer monitor. Then the user feels the interactions with the computer as real. In other words, the *cognitive reality* for a virtual world is established.

Similarly, in our everyday lives, E_U (ego of user) conceptualizes O_R (object in real space) using U (i.e. his self/body) as a reference point. In fact we are not living in a single world but in two worlds at least: our inner world (i.e. our mind) and the outer world

(i.e. the real world). That is the way our cognitive reality for the real world is established.

2.3 Empathy Channel

What becomes a reference point that links our real world and a virtual world on a computer display? We insist that one of the dominant candidates is an empathized ego in the virtual space. As seen in many computer games, a variety of virtual egos of the users exist in monitor, such as a fighter in fighting games and a user's car in driving simulation games. The users can play and step into the game through empathizing with such egos.

On the other hand, we often empathize with protagonists or persons in novels or movies, and virtually experience their lives. Such empathy is achieved by common characteristics between a reader and a character in the story: gender, ethnicity, emotions and experiences in similar situations.

As the empathy connects two cognitive spaces in these occasions and enables the user (or reader) to step into another world, we call it 'Empathy Channel'. The user can acquire and utilize another cognitive viewpoint through the Empathy Channel. Then the user can interact or communicate in the virtual world with a sense of reality. Part of our main goal is to design a good Empathy Channel in human-computer communication environment.

In next section we propose the several methods of enhancing user involvement for movie contents, and analyze TV programs to show how the effective transition of camera shots in them are based on the user involvement to reduce cognitive burden for their audience.

3 Enhancing User Involvement in Movie Contents

In order to enhance the user involvement, human-computer interaction environment should be designed to make the users step into the environment with little effort. Specifically, there should be two kinds of artful devices, that is, establishing a channel to the environment and reducing user's cognitive burden. Thus, it is necessary for making attractive movie-like contents to establish an Empathy Channel in the virtual world and realize smooth shot transition.

In this section we make the following two design suggestions: (1) virtual settings using Empathy Channels for cognitive reality (2) shot transition based on Cognitive Overlapping and Empathy Channel for communicative reality. Furthermore, we analyze popular TV programs and verify the effects of our suggestions.

3.1 Virtual settings with Empathy Channel

As for a conversational agent system, the static aspect is how the virtual world with agents should be arranged. We suggest that the virtual settings using Empathy Channel is one of the effective arrangements.

Nowadays many conversational agent systems are seen as desktop character or web application. Microsoft Agent[1] and Ananova[2] are its typical examples. However, since they work as user's partner, the user will not regard them as his alter ego. As a result, those agents frequently convey the impression that they are just annoyances or unnoticed strangers. There are many reasons for the consequence, but the most important one is that those agents were designed to communicate solely with the users. As observed in human communication, one has to communicate with others via verbal and nonverbal channels. But, natural verbal communication is difficult for conversational agents even if they are 'conversational' because a highly intelligent system is required. On the other hand, it is well-known that one of the nonverbal devices to be established for natural communication is eye contact. However, it is also known that a current agent in a computer display cannot achieve eye contact with its user facing the display. Eventually, the computer users tend to feel those agents as "fake" partners to communicate with.

We thus make a suggestion that a conversational agent should be an empathized ego instead of his partner in two ways. One is to use two or more agents that communicate with each other in the virtual world and to make one agent function as empathized ego of the user. Then the user will empathize with the agent, and will enter the virtual world through it (i.e. Empathy Channel).

The other idea is to use the back image of agent as Empathy Channel (Okamoto et al., 2004). Miyazaki (1993) experimented the empathetic effects of a back image for reading a story. According to the experiment, the picture book featuring the back images of its protagonist make its readers involved in the story more than in the same story with pictures drawn from the observer view. In brief, the back images helped the readers to experience the virtual world as if it were their own.

To sum up, the virtual settings in conversational agent system should be arranged to set Empathy Channels for cognitive reality by applying agent-to-agent communication and back images of the empathized agent.

---

[1] http://www.microsoft.com/msagent/

[2] http://www.ananova.com/video/

3.2 Shot transition with Cognitive Overlapping

In movie contents there are two aspects to be considered for enhancing the user involvement, that is, a static aspect and a dynamic one. Regarding the dynamic aspect of enhancing the user involvement, when making movie-like CG contents that has storytelling progression, smooth shot transition should be realized so as to reduce cognitive burden for the audience.

As many linguistic researches suggest, our discourse and storytelling are based on the consistent flow of new and old information. For example, the following story clearly indicates such information structure:

> *Once upon a time there was a king that wanted a new castle.* ***The king*** *hired the best castle builder in the land to build the castle. Prior to starting* ***to build the castle****, the king got an architect to construct the plans for the castle and* ***the plans*** *were drawn up quickly*[3]*…*

In cognitive linguistics the information structure is considered to be motivated by our cognitive ability of *figure-ground* perception. In the above example, the underlined parts are *figure* while the bold parts become *ground* (Cf. Talmy, 1978; Langacker, 1987).

Nevertheless, not all discourses and stories follow such explicit information flow, but it is certain that the figure-ground alteration would reduce the cognitive burden in reading or listening to a story. We believe that the same structure can be applied to movie contents[4]. See Figure 2 below.

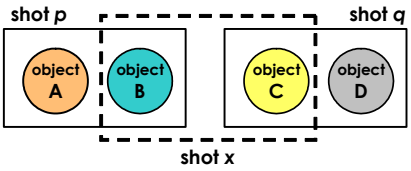

Figure 2: Cognitive Overlapping

In this figure a square corresponds to a camera frame. When shot *p* featuring two objects is just followed by shot *q* featuring other different two objects, the audience is forced to relate the adjacent two shots in his mind, which will be a high cognitive task for him because one movie consists of so many shots that keep changing continuously. In this case, if shot *x*, containing an object already appeared in shot *p*, is inserted between shot *p* and *q*, then the cognitive burden for the audience will be reduced. It is because the object commonly captured in shot *p* and *x* changes its cognitive status from *figure* to *ground* through the shot transition and becomes a reference point for the audience to conceptualize the following shot.

It is assumed that such overlapping is not limited to the visual images of objects. Since movie contents consist of images and sounds, auditory overlapping will also work in enhancing user involvement. For instances, movies or TV dramas sometimes contains the scenes where the shots are changing continuously but narration or actor's voice keeps unchanged during the shot transition.

The overlapping based on figure-ground alteration helps to achieve cognitive reality, so we call it 'Cognitive Overlapping'. The types of shot transition with Cognitive Overlapping are illustrated in Figure 3.

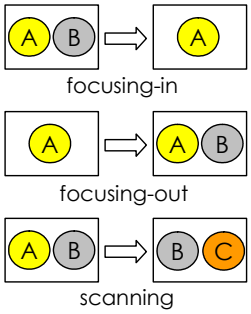

Figure 3: Shot transition with Cognitive Overlapping

Although there are a variety of camerawork techniques such as zoom, tilt, pan, and reverse angle, we classified the shot transition with Cognitive Overlapping into these three types of camerawork. Our classification reflects the semantic relations of shot transition based on captured objects in each shot.

3.3 Shot transition with Empathy Channel

However, we have to admit that non-overlapping transition exists especially from a person shot to an object shot. It seems to contravene our hypothesis, but that is not so.

In such a shot transition type, a person in the previous shot occasionally gives a pointing gesture, gaze, or verbal reference toward the object which lies out of the frame but is recognizable or accessible for him. His attention behaviour leads the audience to make mental contact to the hidden object through the attention as a reference point. As a re-

---

[3] http://www.umsl.edu/~sauter/analysis/fables/fall2002/king_castle.html

[4] Many researchers in film studies have pointed out that language and movie have a lot in common. See Monaco (1977).

sult, the audience gets ready to accept the next shot featuring the object alone (see Figure 4).

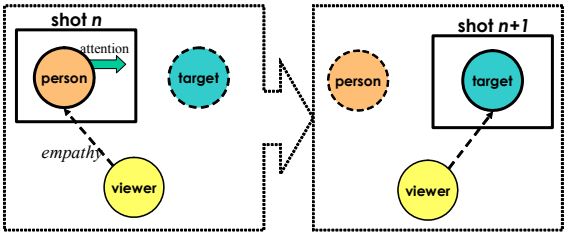

Figure 4: Shot transition with Empathy Channel

In other words, through pointing gesture, gaze or verbal reference by a person in the camera frame, the audience can empathize with him and can effortlessly relate the following shot to the previous one. Such a communicative cue for attention functions as the Empathy Channel for communicative reality.

For establishing a natural human-communication environment, Empathy Channels for both cognitive and communicative reality need to be considered. We thus suggested the virtual settings using Empathy Channels for cognitive reality and the shot transition based on Cognitive Overlapping and Empathy Channel for communicative reality, but our ideas still remain speculative.

3.4 Analysis of the shot transition in TV program

In order to verify the speculation for shot transition, we observed and analyzed certain popular TV programs. This section shows the analyzed data, the analysis, the result and some discussions.

3.4.1 Data

The TV program we analyzed is a 30-minute one providing the information about popular items in a certain shop through the conversation between a TV host as *guide* and a shop owner or clerk as *explainer*. The whole data we used is three programs of 90 minutes.

We divided all the shots in the programs into the following seven types of shots according to what was captured in each shot in view of Cognitive Overlapping (see also Figure 5):

Type 1: The shot featuring the guide
Type 2: The shot featuring the explainer
Type 3: The shot featuring objects to be explained
Type 4: The shot featuring the guide and the explainer
Type 5: The shot featuring the guide and the objects
Type 6: The shot featuring the explainer and the objects
Type 7: The shot featuring the guide, the explainer and the objects

Type 1 Type 3 Type 4

Figure 5: Some examples of shot type

3.4.2 Analysis and results

These programs consist of 485 shots, 78 shots of which is counted in Type 1, 56 shots in Type 2, 117 shots in Type 3, 57 shots in Type 4, 56 shots in Type 5, 54 shots in Type 6, and the other 67 shots in Type 7. As the programs we analyzed were for the information about shops and their items, the most frequent shot type was Type 3, which features an object to be introduced to the viewing audience. Each of the transition rates from one shot to another is shown in Table 1 (Note: the transition between the same shot types is excluded).

Table 1: The shot transition rate (%)

From \ To	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7
Type 1		14.1 (12.8)	17.9 (12.8)	**30.8**	**19.2**	9.0 (3.8)	**9.0**
Type 2	17.9 (16.1)		16.1 (8.9)	**30.4**	1.8 (1.8)	**16.1**	**17.9**
Type 3	17.1	7.7		10.3	**24.8**	**22.2**	**17.9**
Type 4	**26.3**	**40.4**	14.0 (8.8)		**1.8**	**7.0**	**10.5**
Type 5	**33.9**	7.1 (7.1)	**42.9**	**0.0**		**0.0**	**16.1**
Type 6	3.7 (1.9)	**16.7**	**55.6**	**1.9**	**1.9**		**20.4**
Type 7	**6.0**	**4.5**	**58.2**	**6.0**	**13.4**	**11.9**	

In this table, a bold figure means the rate of overlapping transition. The rest represents non-overlapping shot transition. Moreover, each figure in a bracket means the non-overlapping transition occurred with pointing gesture or gaze toward the target in the following shot.

3.4.3 Discussion

This result shows that overlapping shot transition is frequently used in TV programs since the transition occupies 77.9% of the whole transition. It also suggests that it is effective for establishing cognitive reality to use Cognitive Overlapping. Specifically,

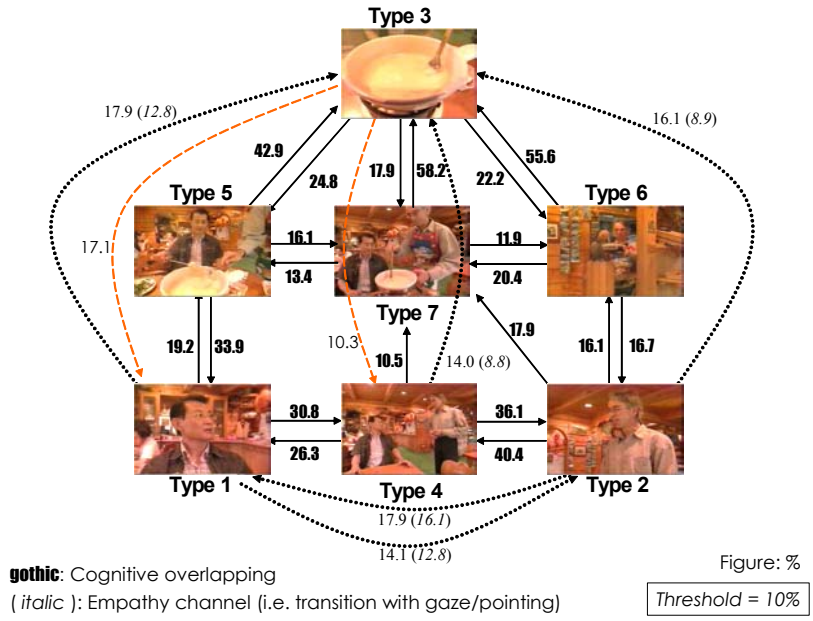

Figure 6: The shot transition model

as illustrated in Figure 3, the *focusing-out* shot transition is used to introduce new information to the audience, while the *focusing-in* to focus the object or the person to pay attention to for the audience.

Although the overlapping transition occupies most part of the whole transition, the rest should be explained in view of the user involvement. We found that 61.6% of the non-overlapping transition is the transition from person shots to object/person shots, that is, the shot transition in which Empathy Channel can be established.

Actually, 72.7% of those transitions include attention behaviours, such as pointing gesture or gaze by the person toward the object(s) in the next shot. Therefore, it can be said that communicative reality via Empathy Channel also enhanced the user involvement. Furthermore, the total rate of the shot transition using either Cognitive Overlapping or Empathy Channel covers as much as 87.8% of the whole.

We summarize the result as the shot transition model in Figure 6. The threshold is fixed at 10%, those transition rates under which are not shown in this model. As seen in Figure 6, there remain a few transitions that would be possible and expected from our theory, but are practically in very low rate or never happen. We assume it might depend on the content or the information flow of the data we used, but more detailed examination is needed.

In next section, we describe our ongoing work for supporting immersive CG contents creation based on the observations here and the virtual settings with Empathy Channel described in Section 3.1.

4 Supporting immersive CG contents creation

In creating CG contents, only skillful artists and creators have so far been able to produce affective and immersive contents with high user involvement. Based on the user involvement theory and the analysis of TV program described in the previous sections, this section proposes a system that enables the user to create immersive CG contents with little effort.

4.1 Previous methods

Virtual Director (Manos et al., 2000) and TVML (Hayashi et al., 1997) are some examples that focus on creating CG contents based on scripting languages. In this approach, all particular events, characters' gestures and also camerawork need to be described in the scripting language beforehand. Thus, creating CG contents using this approach requires sophisticated skill and is quite difficult for amateurs.

In addition, the systems described in Ariyasu et al. (1999) and Douke et al. (2000) try to help the user create CG contents by automatically generating agent arrangement inside a CG set and camerawork. The cost of creating CG contents in such systems can be reduced by using templates, which are derived from the observation of those TV programs that provide information to audience. According to these works, the templates play as tacit rules to enhance the comprehension that the audiences have for each TV program.

Instead of defining templates, in this study, we use the shot transition network proposed in Section 3 as rules for selecting camerawork. This is because the shot transition network can generate more various patterns of camerawork derived from a real TV program than template-based approach. Moreover, our method has an advantage over the previous ones in generating character agents playing on the background photos taken by non-professional users.

4.2 Virtual environment setting

The system we propose is designed for novice users to easily produce movie-like CG contents with enhanced user involvement. Thus, the virtual world setting is simplified and optimized for a specific task, that is, to introduce and explain about interesting objects or monuments to audience by conversation between agents.

The virtual setting of the system is supposed to construct an agent-to-agent communication environment based on Empathy Channel for cognitive reality as illustrated in Figure 7.

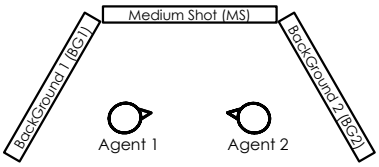

Figure 7: Virtual environment setting

The basic photos that the system uses are one *Medium Shot*, which features an object to be explained, and two *BackGround* photos, each of which becomes a background picture for an agent who stands in front of it.

Additionally, two conversational agents communicate with each other surrounded by the three pictures. One agent is the guide agent, who is expected to function as empathized ego of the user, and the other is the explainer.

For a user to use this system, he needs to prepare at least three photos beforehand, which should satisfy the virtual settings above. Specifically, the photo preparation templates (Figure 8) are available in order to adjust a photo to each template. For instance, a user has to shoot or select a picture which is suitable as background for the placement of two agents to be natural using BG1 template. Similarly, when selecting a photo for MS, the user need to adjust an explanatory object into the dashed square in MS template.

Figure 8: Photo preparation templates

In order to enhance the user involvement, the back image of the guide agent is also used as shown in BG1 template, which generates an over-the-shoulder shot. Then the back image of the guide agent functions as Empathy Channel.

4.3 System architecture

The outline of the system is shown in Figure 9. The system consists of four main modules: the Content Editor, the Shot Generation Module, the Gesture Generation Module, and the Camerawork Generation Module. Using the Content Editor, users select a scene type, three basic photos for BG1, BG2 and MS for the explanatory scene, and speech for animated agents.

Scene type and three photos are sent to the Shot Generation Module, where the photos, the agents and camerawork are properly arranged in the virtual environment. The utterances to be spoken by the agents are sent to the Gesture Generation Module, where agent gestures are selected and scheduled according to the utterances and the scene setting in the virtual environment. Then, in the Camerawork Generation Module, camerawork is specified for each shot based on the shot transition network constructed from the shot transition model in Section 3. The final output of the system is a set of action command executable by the Haptek player[5]. The details of each process are described in the following subsections.

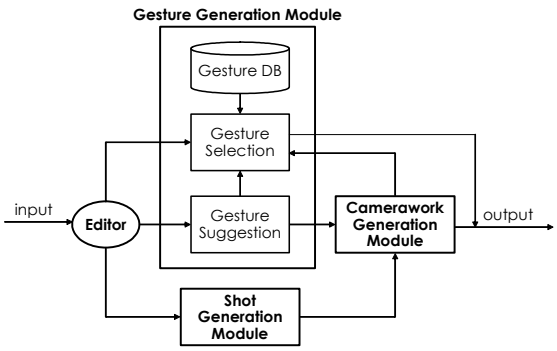

Figure 9: System architecture

[5] http://www.haptek.com/

4.4 Content editor

The users as contents creator specify utterances and photos using the Contents Editor shown in Figure 10.

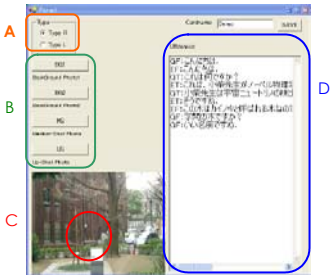

Figure 10: Contents editor interface

The photos and agent utterances are stored as materials constructing a scene. Content editing consists of the following four steps.

First, scene information is specified in (A) in Figure 10. Scene information consists of scene type and placement type. The scene type is either a walk scene or an explanation scene, though our system deals only an explanation scene for the moment.

On the other hand, placement type is either Type R or Type L according to the viewpoint of the guide agent. When the object to be explained (i.e. focused object) is on the right side of the guide agent, the placement type is Type R. When the focused object is supposed to be on the left side of the guide agent, the placement type is Type L.

In picture selection (B), the user selects pictures for the scene. A few photos are selected for a walk scene, and three photos are selected for an explanation scene. BackGround photos (BG1, BG2) are the pictures in which a focused object is not shown. In a medium shot photo (MS), a focused object is shown in a medium distance. There need another photo as UpShot photo (US), which is a zoomed-up picture of a focused object. Therefore, the user can prepare the US separately or might trim the US from the MS, though the US will be of low resolution in that case.

When the system receives the materials, the photos are sent to the Shot Generation Module where all the possible shots are produced from the pictures with character agents. Then the user needs to mark a focused area on the selected MS, which is shown in a preview window (C). Note that a focused object for an explanation scene should not be placed near the edges of the MS photo as described in 4.1.

Utterance information is specified in utterance window (D). First, utterances to be spoken by conversational agents need to be typed in the window. In editing an explanation scene, a speaker tag, which specifies who is the speaker of the utterance, needs to be added to at the beginning of the utterance. A speaker tag "G" is added to the utterances of the guide agent. "E" is added to those of the explainer agent. In addition, "T" tag is added to an utterance which refers to the focused object in it. If the utterance does not refer to the focused object, "F" tag is assigned.

4.5 Shot generation module

In the Shot Generation Module, all the types of shot types defined in Section 3 are generated using setting information in a virtual world and three pictures chosen in the Editor.

Figure 11 shows a virtual environment setting of Type R for an explanation scene. The guide agent (G) and the explainer agent (E) are placed nearly face-to-face. An imaginary line joining the two agents is defined, which properly constrains possible camerawork, and both agents direct 15 degrees away from the imaginary line towards the medium shot. Shot type 1, 2, and 4-7 are produced by camera 1, 2, and 4-7 respectively. Shot type 3 is produced as a zoom-up shot by camera 7.

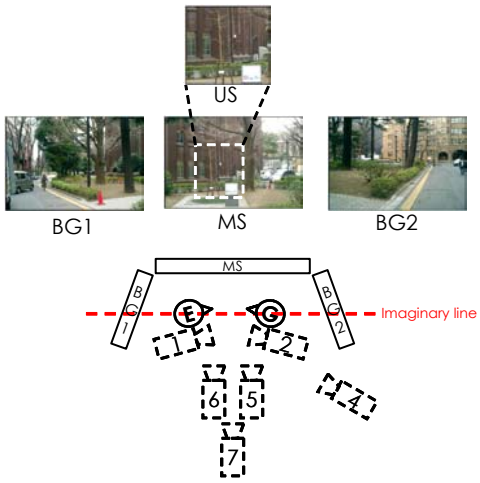

Figure 11: Camerawork generation

In addition, the photos to use for Type 5 and Type 6 are trimmed from the MS as shown in Figure 12.

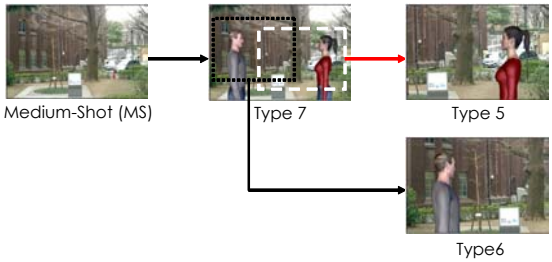

Figure 12: Shot generation from Medium Shot

4.6 Gesture generation

Gesture Generation mechanism consists of two consecutive processes: (1) gesture suggestion in which candidates of gestures are proposed, and (2) gesture selection in which appropriate gesture shapes (e.g., direction of pointing gestures) and gaze direction are determined according to the scene arrangement.

4.6.1 Gesture suggestion

We employ the CAST system (Nakano et al., 2004) as an agent behaviour suggestion mechanism, which outputs the suggestion to the Gesture Selection Module. CAST consists of four main modules: (1) the Agent Behaviour Selection Module (ABS), (2) the Language Tagging Module (LTM), (3) a Text-to-Speech engine (TTS), and (4) a character animation system. When CAST receives a text input, it sends the text to the ABS. The ABS selects appropriate gestures and facial expressions according to linguistic information calculated by the LTM, which uses functions of a Japanese syntactic parser (Kurohashi 1994). Then, the ABS obtains timing information by accessing the TTS, and it calculates a time schedule for the set of agent actions. The output from the ABS is a set of animation instructions that can be interpreted and executed by an animation system. In the proposed system, the timing calculation module is separated from the CAST and used after the gesture selection process.

4.6.2 Gesture Selection

The gesture commands are stored in Gesture Database and are called by the Gesture Selection Module. When the Gesture Selection Module receives suggestions from the Gesture Suggestion Module, it selects appropriate gesture shapes according to a scene setting in the virtual world. At the same time, the Gesture Selection Module receives gaze direction suggestions from Camerawork Generation Module based on the agent placement in the virtual environment setting.

4.7 Camerawork generation

The Camerawork Generation Module produces camerawork for a scene. To produce Cognitive Overlapping and Empathy Channel in CG contents, camerawork for an explanation scene is determined based on the shot transition network shown in Figure 13. Camerawork generation consists of the following three steps.

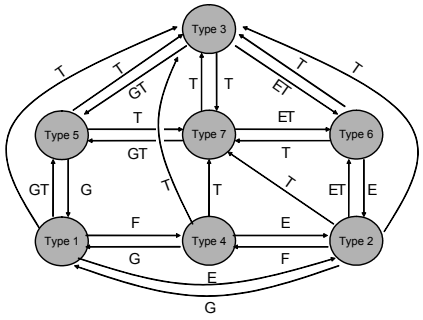

Figure 13: Shot transition network

(1) Determining shot candidates

Shot candidates for each utterance are selected according to who is the next speaker and whether the utterance refers to the focused object or not. Rules for determining shot candidates are shown in Table 2 below.

For example, in the Contents Editor, if the speaker tag "G" (guide agent) and reference tag "T" (referring to a focused object) are assigned to a given utterance, shot type 1, 3, 5, and 7 are selected as shot candidates.

In addition, Camerawork Generation module also manages the time schedule from the Gesture Suggestion Module. For instance, shot is not changed when the duration of an utterance is less than one second. This is because too much shot change makes the contents less comprehensible and increases cognitive burden of the audience. In such a case, a shot type which can be a common candidate for both of the two consecutive utterances is chosen. For example, when utterance A has tag "G" and "F" and utterance B has tag "E" and F", the shot type 4 is continuously used during these two utterances. On the contrary, when the utterance duration is longer than five seconds, a shot is changed not to get the user bored.

Table 2: Shot selection rules

Next speaker	Referring to focused object	Shot type
Guide	F	1, 4
Guide	T	1, 3, 5, 7
Explainer	F	2, 4
Explainer	T	1, 3, 6, 7

(2) Generating shot transition

From the shot type candidates chosen in step (1), this step generates an appropriate shot transition according to the shot transition network. As major transitions of the network produce Cognitive Overlapping, this step generates Cognitive Overlapping camerawork in most of the cases.

(3) Generating eye-gaze

When a transition without Cognitive Overlapping effect is selected, a gaze needs to be generated to produce the shot transition with Empathy Channel, the other device for implementing the User Involvement theory. In this case, a gaze is generated according to the shot transition network and a scene setting given by the user.

5 Conclusion and Future Work

We have so far described our idea on the 'User involvement' as the design theory for establishing natural human-computer interaction environment, and showed our ongoing work on constructing a support system for the immersive CG contents creation, which supports a user who wants to create CG contents with little effort. Since our system has not been fully constructed yet, it still remains unclear what will become obstacles to enhancing the user involvement. However, as long as the system design is based on the framework of the User involvement theory, our system will surely affect and entertain its user. We believe that the virtual settings with Empathy Channel and the shot transition with Cognitive Overlapping and Empathy Channel made the human-computer interaction environment more attractive for the audience of our system.

As our future work, we firstly try to keep constructing our supporting system and then will make a psychological experiment to prove the effect of enhanced user involvement. The experimental results will be shown soon.

Furthermore, since the User involvement theory still remains a design theory that gives a speculative sketch for natural human-computer interaction environment, it is desirable and expected to brush up the theory into that of evaluation.

References

K. Ariyasu, M. Hayashi, H. Sumiyoshi. Automatic Generation of TV Program & Program Relational Contents, *5th Symposium on Intelligent Information Media*, 171-176, 1999.

Mamoru Douke, Masaki Hayashi, Eiji Makino. Automatic Generation of Television News Shows from Given Program Information Using TVML, *Journal of The Institute of Image Information and Television Engineers*, No.7, 1097-1103, 2000.

M. Hayashi, H. Ueda, and T. Kurihara. TVML (TV program Making Language) - Automatic TV Program Generation from Text-based Script, *ACM Multimedia'97 State of the Art Demos*, 1997. http://www.nhk.or.jp/strl/tvml/index.html

R. N. Kraft. The role of cutting in the evaluation and retention of film, *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Vol. 12, No. 1, 155-163, 1986.

Sadao Kurohashi and Makoto Nagao. A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures. *Computational Linugistics*, 20 (4), 507-534, 1994.

Ronald W. Langacker. *Foundations of Cognitive Grammar, vol.1: Theoretical Prerequisites*. Stanford: Stanford University Press, 1987.

Ronald W. Langacker. Reference-Point Constructions. *Cognitive Linguistics* 4: 1-38, 1993.

K. Manos, T. Panayiotopoulos and G. Katsionis. Virtual Director: Visualization of Simple Scenarios, *2nd Hellenic Conference on Artificial Intelligence*, SETN, 2002.

Joseph V. Mascelli, *The five C's of Cinematography: Motion Picture Filming Technique*. Silman-James Press, 1998.

Kiyotaka Miyazaki. The Effects of Human Backimage as Mooring Point in Empathetic Comprehension through Visual Images *(in Japanese). Japan Educational Psychology Association Proceedings*: 35, 1993.

James Monaco. *How to Read a Film: The Art, Technology, Language, History, and Theory of Film and Media*. New York: Oxford University Press, 1978.

Y. Nakano, T. Murayama and T. Nishida. Multimodal Story-based Communication: Integrating a Movie and a Conversational Agent, *IEICE Transactions, Special Issue on Human Communication* (to appear), 2004.

Masashi Okamoto, Yukiko I. Nakano, and Toyoaki Nishida. Toward enhancing user involvement via Empathy Channel in human-computer interface design. In *Proceedings of IMTCI*, 2004.

C.L. Sidner, C. Lee, and N. Lesh. Engagement rules for human-robot collaborative interactions, *IEEE International Conference on Systems, Man & Cybernetics (CSMC),* Vol. 4, 3957-3962, 2003.

Leonard Talmy. Figure and Ground in Complex Sentences. In Joseph H. Greenberg, ed., *Universals of Human Language*, vol.4, *Syntax*, 625-629. Stanford: Stanford University Press, 1978.