Producing effective shot transitions in CG contents based on a cognitive model of User Involvement

Masashi OKAMOTO†, Yukiko I. NAKANO††, Kazunori OKAMOTO†, Ken’ichi MATSUMURA††, Nonmembers and Toyoaki NISHIDA†††, Member

Summary
In virtue of great progress in computer graphics technologies, CG movies have been getting popular. However, cinematography techniques, which contribute to improving the contents’ comprehensibility, need to be learned from professional experiences, and not easily acquired by non-professional people. This paper focuses on film cutting as one of the most important cinematography techniques in conversational scenes, and presents a system that automatically generates shot transitions to improve comprehensibility of CG contents. First, we propose a cognitive model of User Involvement serving as constraints on selecting shot transition. Then, to examine the validity of the model, we analyze shot transitions in TV programs, and based on the analysis, we implement a CG contents creation system. Results of our preliminary evaluation experiment show the effectiveness of the system, specifically in enhancing contents’ comprehensibility.

Key words:
CG movie, cinematography technique, involvement, conversational agent

1. Introduction

As computer graphics technologies have been greatly progressing, they have become indispensable in commercial movie production, and also have been providing tools for non-professional users to enjoy creating short movies. Such CG tools are expected to support the users in creating their own movies and expressing messages and ideas more visually and dynamically than text-based communication (e.g., letter, e-mail, BBS).

However, the richer the media is, the higher the skills required in creating contents are. Videos shot by non-professional people are terribly organized and incomprehensible. Likewise, non-professional CG contents would be worse than text-based messages if CG effects are misused. While filmmakers have cinematography techniques for making the scenes comprehensible for audiences, these techniques are usually learned from experiences, and little has been studied about how to implement such craft knowledge in a computer system.

In this paper, we propose a system that supports non-professional people to create movies comprehensible enough for the purpose of expressing messages and information. Specifically, we focus on conversational scenes where two CG characters are talking about objects in the scene. We are not concerned with artistic cinematography expression, but with enhancing comprehensibility in movie contents by establishing explicit models and rules and implementing them in a system.

In investigating cinematography techniques, we employ a standard framework for describing cinematic structures. A unit of film construction is a shot, which is a single, continuous film recording from the time the camera starts to the time it stops. Individual shots are joined together via a cut, which is a transition between two separate shots [1]. Therefore cutting or shot transition is an important cinematography technique for joining film units (i.e., shots). Psychological research revealed that cutting technique allows the audiences to find consistency between the shots and maintain the narrative structure across the shots despite visual discontinuities in the presented scene [2].

Thus, addressing issues on cutting technique, this paper, based on cognitive linguistics, proposes a model of User Involvement, which derives cognitive constraints on determining shot transitions. This model produces two types of constraints: Cognitive Overlapping and Empathy Channel, both of which contribute to helping the users notice the underlying continuity between the shots. Based on the model of User Involvement, we propose a shot transition model that generates shot transitions reducing user’s cognitive load in understanding a story, specifically in conversational situations. Then, we implement the model in a movie creation support system.

In the rest of this paper, first, we review previous related work in section 2. In section 3, we propose a model of User Involvement, and, based on the model, discuss the characteristics of shot transitions that reduce user’s cognitive load in understanding. Then, to examine the validity of the proposed model, we analyze shot transitions in TV programs in section 4. Based on the data analysis, a CG contents creation system is presented in section 5, and results of a preliminary evaluation experiment are reported in section 6.
2. Related Work

2.1 Cinematography studies

Although it is well known that there are many cinematography techniques which are frequently used by professional movie creators, little is known about how to use each technique on making movie contents.

Among those techniques, Kraft [1] suggests that ‘cutting’ has two psychological functions: a syntactic function by segmenting the flow of filmed activities and a rhetorical function by influencing the connotative and affective characteristics of film sequences. Through examining the cutting effect, he found that films with cutting gave more interesting, active and stronger impressions on the viewers though the viewers hardly remember the number of cuts. These findings suggest that effective cuttings will enhance the viewers’ involvement with movie contents while inappropriate cuttings will disturb their concentration. However, it is unclear how to build an effective cutting model for movie creation.

On the other hand, May et al. [3] also focuses on the cutting effects and examined the correlation between cuts and viewers’ cognition. Based on their own cognitive model[4], they put general hypotheses, verified them through an experiment, and found that most of cuts have strong relationships with the cognition of the viewers. For example, eyeline axis (i.e. imaginary line) in a conversational scene is very important for the viewers to comprehend the scene. In addition, what is captured in each shot affects the viewers’ cognitive burden. Therefore, considering the viewers’ cognition is indispensable for making movie contents.

2.2 CG contents creation

Virtual Director [5] and TVML [6] are some examples that focus on creating CG contents based on scripting languages. In this approach, all particular events, characters, gestures and camerawork need to be described in the scripting language beforehand. Thus, creating CG contents using this approach requires expert skill and is quite difficult for amateurs.

In addition, the systems described in [7, 8] 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 model proposed in Section 4 as rules for selecting camerawork. This is because the shot transition model can generate more various patterns of camerawork derived from real TV programs than the 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.

3. User Involvement for CG movie contents

In this paper, we apply the ‘User Involvement’ theory [9] to the designing of CG movie as natural human-computer interaction environment. Our research focuses on finding out how the User Involvement can be established and enhanced through smooth shot transition and effective virtual settings in CG movie contents.

3.1 Establishing User Involvement

The User Involvement theory was conceived to design a natural human-computer interaction environment in which humans willingly engage or forced to be involved[9]. According to the theory, the computer user should feel either cognitive or communicative reality for the interaction or communication with a computer. Based on the cognitive linguistics studies, especially the reference-point ability [10], it is suggested that a user can conceptualize an entity in a cognitive space by way of a certain reference point which connects two cognitive spaces.

![Figure 1: Astigmatic model of User Involvement](image)

Figure 1 illustrates an example of how the computer user conceptualizes objects in different cognitive spaces using each reference point, especially in virtual agent systems. We call this model the astigmatic model for User Involvement in that multiple spaces are overlapped and linked there. For example, $E_U$ (ego of user) locates $O_U$ (object in real space) using $U$ (i.e. his self/body) as a reference point. It is one of our everyday activities in which we conceptualize something in the real world. On the other hand, $U$ can conceptualize $O_A$ (object in virtual space) via a certain reference point which is cognitively
accessible for the user, and, at the same time, consistently functions as a constituent of the virtual world.

3.2 Enhancing User Involvement in Movie-like Contents

In order to enhance the User Involvement, human-computer interaction environment should be designed to make the users involved in the environment with little effort. Specifically, there should be two kinds of artful devices, one of which is establishing a channel to the environment, and the other reducing user’s cognitive burden.

We thus make the following two design suggestions for creating CG movie contents: (1) virtual environment setting using Empathy Channels for cognitive reality, (2) shot transitions based on Cognitive Overlapping and Empathy Channel for communicative reality.

3.2.1 Virtual environment setting with Empathy Channel

Nowadays many conversational agent systems are seen as desktop character or web application. Microsoft Agent [11] and Ananova [12] are its typical examples. However, since they work as user’s partners, 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 uninvited visitors. The most important reason for this consequence is that these agents do not have awareness of their surroundings.

To design a conversational agent that can behave as a genuine partner of a user, the agent should be able to communicate with the user via verbal and nonverbal channels. Nevertheless, natural verbal communication is still difficult for conversational agents even if they are ‘conversational’, because a highly intelligent system is required but is not good enough for practical use. 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 appropriate eye contact with its user facing the display.

We thus claim that a conversational agent should be an empathized ego instead of playing a user partner’s role. In order to realize an empathized agent, we propose the following two methods.

(a) Empathized agent: One is to build a virtual environment where two or more agents communicate with each other and one of them should serve as an empathized ego of a user. Empathizing with the agent, the user can enter the virtual world through it. In other words, the empathy here functions as a reference point which connects the real and the virtual world. We thus call it ‘Empathy Channel’, through which the user acquires and utilizes another cognitive viewpoint. The user is able to interact or communicate in the virtual world with reality.

(b) Back image: Using the back image of an empathized agent effectively facilitates Empathy Channel. Miyazaki [13] assessed the empathetic effects of a back image for reading a story. According to his experiment, the picture book featuring back images of its protagonist makes its readers more involved in the story than in the same story featuring pictures drawn from the observer’s point of view. In brief, the back images helped the readers to experience the virtual world as if it were their own.

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

3.2.2 Shot transition with Cognitive Overlapping

In CG movie contents, there is another aspect to be considered for enhancing the User Involvement, that is, a dynamic aspect. As for the dynamic aspect, smooth shot transition should be realized for the purpose of reducing cognitive load for the audience.

As many linguistic researches suggest, our discourse and storytelling are based on a consistent flow of new and old information. For example, the following story clearly indicates such an 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...*[14]

In cognitive linguistics, the information structure of discourse is considered to be motivated by cognitive ability of *figure-ground* perception. In the example above, the underlined parts are *figure* while the bold parts become *ground* [14].

Needless to say, 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.

As shown in Figure 2, a square corresponds to a camera frame. When shot $p$ featuring two objects is just

![Figure 2: Cognitive Overlapping](image-url)
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. It is 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 \( q \), and then the cognitive burden for the audience will be reduced. The reason is that the object commonly captured in shot \( p \) and \( x \) changes its cognitive status from \textit{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 visual images of objects. Since movie contents consist of images and sounds, auditory overlapping may also work in enhancing User Involvement. For instance, movies or TV dramas sometimes contain 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](image)

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

### 3.2.3 Shot transition with Empathy Channel

However, we have to admit that there are non-overlapping transitions 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, a person in the previous shot occasionally gives a pointing gesture, gaze, or verbal reference toward an object which lies out of the frame but is recognizable or accessible for him. His attention behavior leads the audience to make mental contact to the hidden object through the attention as a reference point. As a result, the audience gets ready for the next shot featuring the object alone (see Figure 4).

![Figure 4: Shot transition with Empathy Channel](image)

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

To achieve a natural communication between humans and computers, Empathy Channels for both cognitive and communicative reality need to be considered. We thus suggested (1) the virtual environment setting using Empathy Channels for cognitive reality and (2) the shot transition based on Cognitive Overlapping and Empathy Channel for communicative reality. We will conduct an empirical study to examine the validity of our model.

### 4. Data Analysis

#### 4.1 Data

In order to verify our model of shot transition, we observed and analyzed three 30-minute TV programs. They provide information about popular items in a certain shop through the conversation between a TV host as guide and a shop owner as explainer.

We classified all the shots in the programs into the following seven shot types according to what was captured in each shot with respect to Cognitive Overlapping:

**Shot types**
- 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
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 those 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 modeled in Figure 5 (Note: the transitions between the same shot types and those below 10% occurrence are excluded).

![Figure 5: Shot transition model](image)

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 Cognitive Overlapping is effective for establishing cognitive reality. Specifically, as illustrated in Figure 5, the focusing-out shot transition is used to introduce new information to the audience, while the focusing-in to notify the audience of what s/he should pay attention to.

Although the overlapping transition occupies a large part of the whole transition, the rest should be explained with respect to the User Involvement. We found that 61.6% of the non-overlapping transitions are transitions from person shots to object/person shots, that is, the shot transition in which Empathy Channel is possibly established. Actually, 72.7% of those transitions include attention behaviors, such as a pointing gesture or a gaze toward the object(s) in the next shot. Therefore, this suggests that communicative reality via Empathy Channel also enhanced the User Involvement. Furthermore, the total rate of the shot transitions using either Cognitive Overlapping or Empathy Channel covers as much as 87.8% of the whole.

As seen in Figure 5, there are a few transitions that are predicted by our model but not happened in real data. This consequence might depend on specific content or information flow of the data we used, but more detailed examination is needed.

In next section, we describe our system for supporting immersive CG contents creation based on the observations here and the virtual settings with Empathy Channel described in Section 3.

5. Supporting CG movie contents creation

In creating CG movies, 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 programs described in the previous sections, this section proposes a system that enables the user to create CG movie contents with little effort.

5.1 System architecture

The outline of the system is shown in Figure 6. The system consists of four main modules: the Content Editor, the Shot Generation Module, the Gesture Generation Module, and the Camerawork Generation Module. Through the Content Editor, the users specify scene information, photos served as background of a CG content, and utterances spoken by character agents. Then, in the Shot Generation Module, photos, agents and camerawork are properly arranged in a scene. The utterances are sent to the Gesture Generation Module, where agent gestures are selected according to the scene setting in a virtual environment, and also scheduled in a way of synchronizing with speech. Then, in the Camerawork Generation Module, camerawork is specified for each shot based on the shot transition model described in Section 3. The final output of the system is a set of action commands.
executable by the Haptex player [16]. The details of each process are described in the following subsections.

5.2 Content editor

The users as contents creator specify utterances and photos using the Content Editor shown in Figure 7. The photos and agent utterances are stored as materials constructing a scene. Content editing consists of the following four steps.

**Specifying placement type:** First, placement type is specified in A in Figure 7. 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.

**Selecting pictures:** In B, the user selects pictures for the scene. In order to construct an agent-to-agent communication environment facilitating Empathy Channel, we propose a virtual environment setting (VES) simplified and optimized for narrative explanations performed by two conversational agents as shown in Figure 8. The user needs to prepare at least three photos beforehand, which should satisfy the VES. The basic photos are one Medium Shot (MS), in which a focused object is shown in a medium distance, and two Background photos (BG1 and BG2), each of which becomes a background picture for agents talking in the scene. In addition, UpShot photo (US), which is a zoomed-in picture of a focused object, can be trimmed from the MS.

As shown in Figure 8, in the VES, two conversational agents communicate with each other surrounded by the three pictures. One agent is the guide agent serving as an empathized ego of the user, and the other is the explainer. To enhance the User Involvement, specifically Empathy Channel, the system generates an over-the-shoulder shot by using a back image of the guide agent on the BG1. 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.

**Marking a focused area:** In a preview window C, a focused area on the MS needs to be marked. Note that a focused object for an explanation scene should not be placed near the edges of the MS photo.

**Utterance information:** Agent utterances need to be typed in utterance window D, and the speaker of the utterance needs to be specified by annotating a speaker tag: “G” for utterances of the guide agent, and “E” for those of the explainer agent. In addition, if an utterance refers to a focused object, “T” tag is annotated to the utterance. If not, “F” tag is assigned.

5.3 Shot generation module

In the Shot Generation Module, all the shot types defined in Section 4 are generated using the VES and three pictures chosen in the Content Editor. Figure 8 shows the VES for a Type R scene. The guide agent (G) and the explainer agent (E) are placed nearly face-to-face. The two agents are joined with an imaginary line, which serves as a constraint on selecting possible camerawork. Both agents direct 15 degrees away from the imaginary line towards the MS. 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-in shot from camera 7. In addition, shot types 5 and 6 are trimmed from the MS as shown in Figure 9.

![Figure 7: Content editor interface](image)

**Figure 7:** Content editor interface

![Figure 8: Virtual environment setting and camera positions](image)

**Figure 8:** Virtual environment setting and camera positions

![Figure 9: Shot generation from Medium Shot](image)

**Figure 9:** Shot generation from Medium Shot
5.4 Gesture generation

Gesture Generation mechanism consists of two consecutive processes: (1) suggesting candidate gestures, and (2) selecting appropriate gesture shapes (e.g., direction of pointing gestures) and gaze direction according to the scene arrangement.

**Gesture suggestion**

We employ the CAST system [17] as an agent Gesture Suggestion Module (GSGM), which outputs candidate behaviors to the Gesture Selection Module (GSLM). We use three modules in CAST: (1) the Agent Behavior Selection Module (ABS), (2) the Language Tagging Module (LTM), and (3) a Text-to-Speech engine (TTS). 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[18]. Then, the ABS obtains timing information by accessing the TTS, and 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.

**Gesture Selection**

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

5.5 Camerawork generation

The Camerawork Generation Module (CGM) produces camerawork for a scene. To produce Cognitive Overlapping and Empathy Channel, camerawork for an explanation scene is determined based on the shot transition model shown in Figure 5. Camerawork generation consists of the following three steps.

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

In addition, the CGM manages the time schedule sent from the GSGM. When the duration of an utterance is less than one second, the next shot change is prevented by the CGM. This is because too much shot change makes the contents less comprehensible and increases cognitive burden on the audience. In such a case, a shot type which is applicable to 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 preferred to be changed not to get the user bored.

<table>
<thead>
<tr>
<th>Next speaker</th>
<th>Referring to focused object</th>
<th>Shot type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guide (G)</td>
<td>F</td>
<td>1, 4</td>
</tr>
<tr>
<td>Guide (G)</td>
<td>T</td>
<td>1, 3, 5, 7</td>
</tr>
<tr>
<td>Explainer (E)</td>
<td>F</td>
<td>2, 4</td>
</tr>
<tr>
<td>Explainer (E)</td>
<td>T</td>
<td>1, 3, 6, 7</td>
</tr>
</tbody>
</table>

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

(3) **Generating eye-gaze:** When a selected transition lacks Cognitive Overlapping, Empathy Channel, which is the other device of User Involvement, is produced using eye-gaze behavior by the speaker agent. The gaze direction is calculated according to a scene setting specified in the Content Editor.

6. Preliminary evaluation

In order to examine whether the CG contents created by our system are more comprehensible than those without shot transition, a preliminary evaluation experiment was conducted.

**Procedure:** Twelve subjects (eight women and four men) answered a questionnaire asking about their impression on watching CG contents. Each subject watched two different CG stories. If a subject watched Story A with shot transitions produced by our system, s/he watched Story B without shot transition, and vice versa. The items (i.e., questions) in the questionnaire were classified into two types: (a) questions about general impression on the contents, (b) those about comprehensibility of the contents. Type (a) questions asked how expressive and less boring each movie is. Type (b) questions ask whether it is easy to understand the story and grasp the point.
**Results:** Average scores for each question type are shown in Figure 10. Average score for general impression (measured by type (a) questions) on the proposed method is higher than that on without shot transition (3.17 and 2.58 respectively), but the result was not statistically significant (t(46)=2.013 (n.s.)). On the other hand, effects of shot transition measured by type (b) questions are statistically significant (t(94)=3.215, p=0.002) Average score for contents with shot transitions is 4.19, and that for those without transitions is 3.40.

In order to confirm whether the subjects actually remembered important points in the story when the contents were presented with shot transitions, we asked the subjects to recall items shown in the movies as much as possible. As a result, four of six subjects watching the same contents remembered a key object in a story when the object was presented with shot transitions. On the other hand, only one subject remembered the focused object when it was presented without shot transitions. While these results are still preliminary, they suggest that user involvement is one of the major factors affecting comprehensibility of movie contents.

![Figure 10: Results of preliminary evaluation.](image)

**7. Conclusion**

We have so far described our idea on the ‘User Involvement’ as the design theory for establishing natural human-computer interaction environment. Moreover, we proposed a cognitive model for CG movie contents, in which the shot transition and the virtual environment setting based on Cognitive Overlapping and Empathy Channel, contribute to enhancing the User Involvement. Based on the model, we presented a support system for CG contents creation, which supports non-professional users to create CG movies with little effort. Though we admit that our system needs to be improved, the system surely enhances the comprehensibility of the movies as long as the system design is based on our cognitive model of the User Involvement.

Since the User Involvement theory 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**