Journal of VLSI Signal Processing 27, 141–160, 2001

Concept and Construction Example of an Interactive Movie System

RYOHEI NAKATSU, NAOKO TOSA AND TAKESHI OCHI
ATR Media Integration & Communications Research Laboratories, 2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan

Received August 2, 1999; Revised January 19, 2000

Abstract. Interactive movies are proposed as a new type of media produced by combining new trends found in movies, communications, and games. With interactive movies, we are able to build cyberspaces rich to the feeling of presence by applying computer graphics and 3D observation, and to get the feeling of complete immersion in these spaces, as if we were actually in them. In addition, it is possible to experience stories through interaction with other characters in the spaces. We first explain the concept of interactive movies and describe a first prototype system that we have developed. We then describe the construction of a second system, which we are currently developing, as well as several improvements in the system.

Keywords: interactive movies, new media, interactions, autonomous characters, speech recognition, emotion recognition, gesture recognition, computer graphics

1. Introduction

Recently, significant common trends in the worlds of communications and entertainment have been appearing, such as communications in cyberspaces and entertainment in multiuser virtual worlds. For example, in the world of communications, many applications for end user participation have been developed on the Internet. The Internet can be thought of as a huge cyberspace connecting people from all over the world [1]. In this cyberspace, people can communicate with others and are also able to form communities of interest. As another example, in the field of movies, recent movies have incorporated digital technologies and computer graphics technologies that are evolving towards a new generation of movies. Digital and computer graphics technologies provide us with very life-like worlds not seen in movies to date; in other words, they give us the ability to create cyberspaces [2, 3]. In addition, video games, especially role playing games (RPGs), make it possible for us to enjoy a story as the main figure in a cyberspace. One possible explanation for the popularity of RPGs is that the story experiences that we get in cyberspaces reinforce our natural tendency to engage in creative story telling [4, 5].

The creation of a new media form, which incorporates elements from emerging trends in communications, movies, and games, as mentioned above, is the current focus of many research programs. As one of such media, there is a concept of interactive movies. Although there have been many researches and attempts on interactive movies, still its concept is quite vague and not ample. In this paper, as one of the many possible forms, we propose a new type of interactive movies. Our interactive movies produce cyberspaces rich to the feeling of presence by the application of computer graphics, actual photos, and 3D observation, and give people the feeling as if they were actually in the spaces. In addition, they make it possible for us to have interactions with characters in the cyberspaces. From this, it becomes possible for a person to experience a story by becoming the main figure in a cyberspace.

Based on this viewpoint, we have been conducting research on interactive movie production by applying interaction technologies to conventional movie making techniques. As an initial step in creating this

new type of movie, we have produced a prototype system [6]. By evaluating this system, we have learned that spontaneous interaction is the key element for subject participation in narratives. Based on this evaluation, we are currently developing a second prototype system with many improvements [7].

In this paper, we explain the basic concept of our interactive movies having the above possibilities, conditions for system construction, and a concrete system example. First, in Section 2, the concept of interactive movies is given and a comparison is made with other types of media. Section 3 examines the conditions when constructing an interactive movie system. Then, Section 4 explains the contents of our first prototype system. Lastly, Section 5 introduces the configuration of our second prototype system, which is now being refined by incorporating the described improvements.

2. Positioning of Interactive Movies

2.1. Concept

When we think about story experience in a cyberspace, we tend to have some fixed image, especially when the thinking involves video games and virtual theaters in theme parks, but we do lack knowledge on the means of existence as examples are limited (i.e., no systematic explanations). Recently, J.H. Murray discussed this, and explained the possibility of the emergence of new tales that exceed the storytelling of past novels [4]. However, her points of argument remain at an abstract level, and have not led to the proposals of concrete systems carrying out storytelling in cyberspaces. Consequently, there are different types of approaches such as approaches based on novels, movies, or games.

To address this, we aim at the construction of a system able to give story experiences in cyberspaces, with movies [8] being the starting point. Our goal is to establish a position for such movies as a new type of media incorporating various types of storytelling techniques that have been devised over 100 years since their birth. There have been many attempts to introduce an interactive capability into movies. However, in most of these attempts, as will be described later, the role of the audience remains as only viewers, not participants. There is another type of approach where the members of the audience are not mere viewers of the movie. Instead, a person can participate as a hero or heroine by interacting directly with other characters in the movie and can therefore experience the story as it proceeds. As one approach we propose “Interactive Movies” in this paper (To identify the interactive movies proposed here from the general concept of interactive movies, we use the term “Interactive Movies” in this paper when necessary.)

First, we will give a simple definition of our Interactive Moves. Interactive Movies in practical use should allow participants to experience stories as main characters by using an Interactive Movie System. The following functions comprise our Interactive Movie System.

(1) A function to create an interactive story where the ongoing progress and ending change based on participant interaction. Such a story is constructed as a series of scenes.
(2) A function to produce scenes constructed with images (which are projected or displayed on a screen or display) and with voices and sounds related to the images.
(3) A function to produce and control movements of singular or multiple characters performing roles other than the main character in a story.
(4) An interactive function to make interaction possible between participants and other characters or objects in scenes using natural interfaces such as voices or gestures.

In [4], an explanation is given that the necessary characteristics of a story experience system in a cyberspace are immersion and interactivity. In our Interactive Movies, the following points are characterized more concretely.

(1) *Structure of the cyberspace and the immersion of the participants in the cyberspace*: By integrating the practical uses of computer graphics and actual photo images, and moreover, by creating them in 3D, we construct a cyberspace rich to the feeling of presence and give the participants a feeling of actually being in the space.
(2) *Story experience in the cyberspace involving interaction*: Storytelling can be experienced when the participants interact with the various characters in the cyberspace through the use of voices, body movements, and hand movements.

2.2. Comparison with Other Types of Media

Here, we compare Interactive Movies with other types of media.

(1) *Communications*: By reproducing people at distant locations and their surrounding environments with 3D images, research is progressing to enable communications rich to the feeling of presence, somewhat like communicating face-to-face. A representative example is the presence-inducing teleconferencing research [9] being carried out at ATR. This research involves telecommunications introducing a cyberspace, but does not include concepts of stories, i.e., merely progressing with the advancement of teleconferencing.

(2) *Movies*: Movies have come to draw humans into the world of fiction (i.e., cyberspaces) by powerful images and sounds, and by appealing to their sense of sight and their sense of hearing [4, 8]. In particular, recently, it has become possible to produce happenings (that are in fact improbable) as realistic images in the world of fiction by using the recent computer graphics technologies. The idea of adding the function of interaction to such movies has been considered. Some of the early attempts prepared a number of storytelling situations and showed one of them to the members of an audience depending on their demands. Recently, with advances in computer technologies, it has become possible to change the scenes or stories of movies in real time. This has led to the idea of new movie types where an audience can select one of several story lines or select one of several viewpoints based on computer control. Here the important issue is how to present smooth and seamless story lines regardless of the choice [10]. In almost all of these attempts the role of the audience remains as a viewer and not as a participant, in spite of the fact that there is a unique approach to introduce cinematic and dramatic techniques into computer based interactive performances [11].

(3) *Video games*: Video games, especially role playing games (RPGs), make it possible to tailor the world of novels to games. Basic stories are set, and humans can control the storytelling by manipulating the main figures in the games. In this light, video games look as if to be very close to Interactive Movies, but interaction is performed by the pressing of buttons. On the other hand, for Interactive Movies natural interaction is introduced through the usage of voices and gestures of humans themselves. This natural interaction capability is the basis for creating immersion.

There is also a big difference between the above two media from the viewpoint of story construction. For video games, the intent of most stories is to get rid of the enemy and to attain the goal. On the other hand, for Interactive Movies, the story tries to describe inspired human hearts much like the main flow of novels and movies [4].

(4) *Other types of media*: There have been many trials to create computer characters that can communicate with humans. The most successful example among them is ELIZA [12], which achieves conversation based on typed-in messages. The basic conversation capability of ELIZA stands on a simple echo back algorithm and has been successful in very limited areas such as a conversation between a patient and a therapist.

Recently, various ongoing research works have been trying to build cyberspaces, and to create interaction in these worlds between characters with visitors. This has included the creation of computer characters [13–17] able to interact at the performance level and emotional level of humans, and interactive art [18]. However, these types of interactions are short-duration events, and no story is introduced. Another research work has shown that, by introducing reactive autonomous computer characters, an audience can play the performance in collaboration with computer characters [19]. In the sense that the audience becomes the main character, this resembles our Interactive Movies. However, the performance emphasizes the experience of spontaneous interactions, and little attention is paid on how to treat stories and what are the important issues in story experience.

Interactive Movies, where storytelling is basically cinematic in nature, aim at proceeding to take an interaction function. Table 1 shows, with [4, 5] as reference, a detailed comparison of movies and video games, among the above-mentioned types of media, and Interactive Movies.

3. Conditions for Creating Interactive Movies

There are several basic elements of Interactive Movies, as explained in 2.1. Here, we consider very simply some of the conditions that should be including these functions.

Table 1. Comparison of movies, video games, and interactive movies.

	Objective	Audience/Participant	Story	Interaction
Movies	Allows audience to experience emotional satisfaction through experience of a story.	Observers of story development.	Key factor deciding quality	None
Video games	Allows players to experience aggressove roles.	Subject of games.	Supplementary factor	Button input
Interactive movies	Allows participants to experience positive and immersive experiences through participation in story	Subject experiences stories Difference from movies: subjective experience. Difference from games: Mental satisfaction.	Key factor deciding quality	Voice and gesture interactions that are natural means of human communication.

3.1. Story Creation

The following types of stories can be considered for interactive storytelling.

(1) *Semi-free storytelling*: Semi-free means that the outlines of the stories are fixed beforehand. However, some degree of freedom (or some variety in storytelling) is set midway, and participants can enjoy the freedom of storytelling with interaction introduced within their surroundings. Almost all RPGs at present can be thought of as being of this type. An advantage of this type is the possibility of fabricating the storytelling beforehand, with the know-how amassed in the making of novels and movies of the past. By doing this, participants can become totally absorbed into the story world.

(2) *Full-free storytelling*: Full-free storytelling is where the story is not decided beforehand, and it is said to involve ongoing development depending on the interaction of the participants. In this case, the production side constructs the world where the story is to develop, and later, everything is left up to the free will of the participants. In this method, it is important to proceed with development by considering all possibilities, because the actions of the participants cannot be predicted beforehand. Even from the "play" side, to be given complete freedom is not always advantageous in that participants may not know how to behave. Although full-free storytelling is the ultimate target of interactive storytelling, as indicated above, this is very difficult. A good example of this type of storytelling is the Oz project [15]. In this project they tried to construct a simulated world inhabited by fully autonomous agents and, by interacting with them, people can participate in the world. Within a limited area they have succeeded. However, the Oz world is not as complex as the real world and does not give people deep experiences.

As we want to let people have deep emotional experiences based on sophisticated stories, we have adopted semi-free storytelling for our Interactive Movies.

3.2. Character Creation

Reaction patterns of computer graphics characters and their movement sequences are prepared by a method that prepares everything beforehand as animation and a method that automatically makes movements and reactions of a certain degree. The ultimate goal is to make characters be autonomous. Accordingly, research is quite active on autonomous characters [13, 15–17, 20, 21], but it is fact that we have not yet reached the point where there is a sufficient amount of autonomy. Consequently, while animation creation is based on an animator, an approach is desired able to add autonomy little by little to such animation.

3.3. Interaction

Communications among humans themselves proceeds through multi-modal interaction. Accordingly, it is hoped that not only one type of modality (such as movements or voices) but a number of them such as movements and voices will be used. This conceivably can lead to the realization of more natural interaction, and on account of that, the introduction of immersion and emotions at deep levels.

In addition, the transmitting and receiving of information on emotions and sensitivity play an important

role in our communications [22]. The authors have therefore come to do research on creating characters having a function to respond appropriately after recognizing emotions included in the voices of humans [14]. What we have learned from exhibitions is that humans become immersed in their interaction with these characters by overcoming the barriers of sex, age, and moreover, language. From this experience, it can be thought that communications through emotions is an important element of interactive movies.

4. System Construction of the First Prototype System

Based on the above considerations, we constructed a first prototype of an interactive movie system. We explain the concrete details of the system below.

4.1. Characteristics

(1) *Image expressions with a feeling of presence*: We make our environmental settings rich to the feeling of presence by projecting 3D images onto an arch-shaped screen, and aim to draw visitors into the world of the interactive movie.
(2) *Natural interaction*: By using speech recognition and gesture recognition functions, we achieve interaction by the use of voices and gestures deemed natural by humans. Moreover, the communications include emotions.
(3) *Computer graphics animation for multi-stories*: For interactive stories, there is a need to prepare large animations because of the complexity of storytelling. It is possible to avoid problems by making the computer graphics characters autonomous, but the naturalness of the animations is lost. Considering all of this, we prioritize the degree of completion for animations, and take an approach that prepares all important animations in advance. In addition, one of the authors, who is an artist, creates the animations and makes considerations in preparation for character reactions and movements that are natural and human-like.

4.2. Software Construction

Figure 1 shows the software construction.

(1) *Script manager*: This part creates concrete interactive scenarios from scripts and scenarios made by story writers and controls all parts of the storytelling of interactive stories. Interactive stories are expressed with a merging of various scenes and conditional change figures between the scenes. In addition, each scene is expressed with a conditional change figure between two various shots (Fig. 2).

The script manager memorizes these conditional change figures and with the interaction results sent from the scene manager, controls the changes between scenes or shots.
(2) *Scene manager*: The content descriptions of individual scenes are stored as data in advance. The scene manager references the description data directed from the script manager and then creates individual scenes. Figure 3 shows the description form of scenes. It is constructed from the following

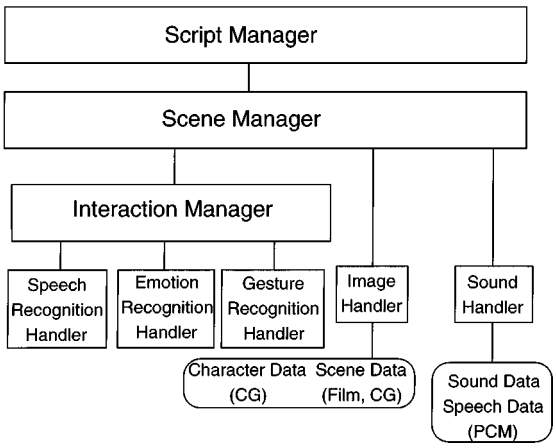


Figure 1. Software configuration of the Interactive Movie system.

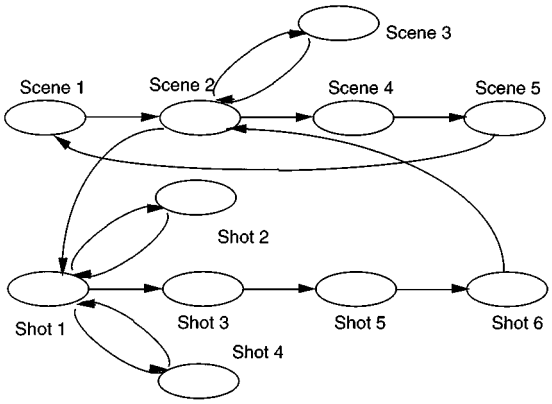


Figure 2. Scene and shot transition diagram.

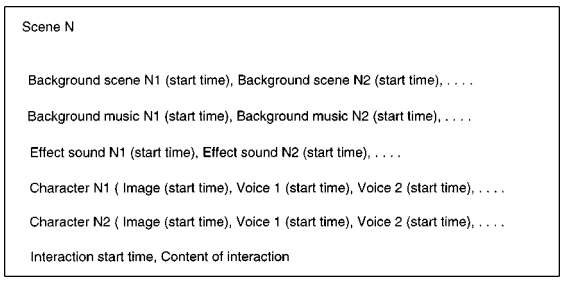


Figure 3. Scene data format.

elements.

a) Background scene and background sound, and their starting times
b) Effect sound and the starting time of its output
c) Animation of the characters and the starting time. In addition, the lines of the characters and the starting times of their output
d) Types of interactions between participants and characters, and the starting times

Background images are constructed from mixtures of computer graphics images and actual images. Following the opening of each scene, the scene manager sends commands to the image handler and the sound handler, and starts the output of the background image and the output of the background sound. The characters are produced from computer graphics. Based on the information in (c), the scene manager sends a command to the image handler and starts the animation of the characters at the starting time. In addition, it sends a command to the sound handler and performs output of the applicable lines at the starting times of the lines. When multiple characters appear, it performs this processing for each of the characters. Moreover, at the starting time of the interaction, it sends a command to the interaction manager and makes the participants and characters start to interact. When it receives the interaction results from the interaction manager, it sends the results to the script manager, and then prepares to create the next scene. Figure 4 shows the time sequence of the processing of the scene manager.

(3) *Interaction manager*: The interaction manager exists below the script manager and the scene manager, and controls the interaction in each scene. Interaction is done based on a speech recognition function, emotion recognition function, and gesture recognition function. When it receives types of interactions and starting commands from the scene manager, it sends commands to the appropriate handlers, and makes either the speech recognition function, emotions recognition function, or gesture recognition function start or makes various types of movements start. When it receives the recognition results, it creates the final interaction

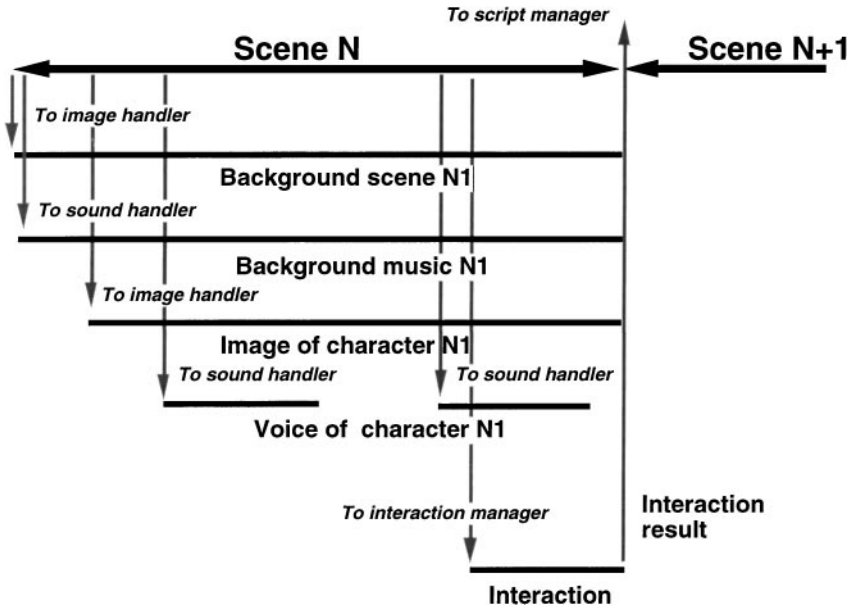


Figure 4. Time sequence of scene manager processing.

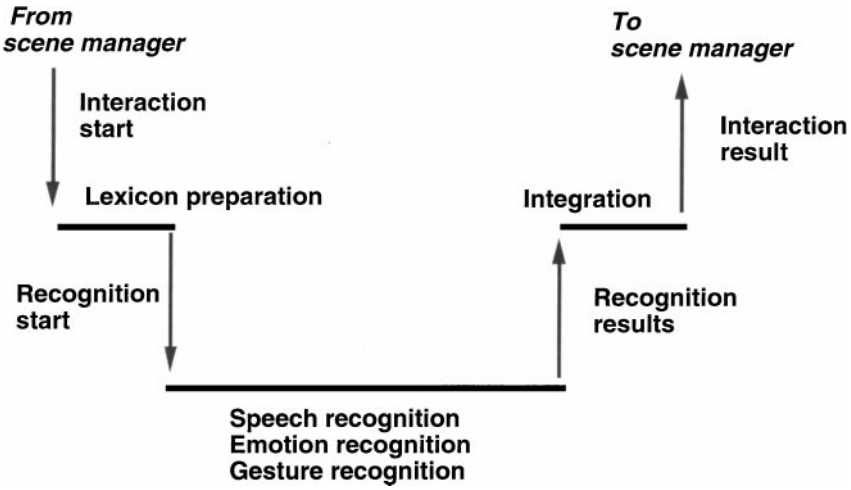


Figure 5. Time sequence of interaction manager processing.

results by combining the recognition results, and sends the information to the scene manager. Concerning how many types of recognition results are combined, because this is deeply related to the contents of the individual stories, the method is decided for each story. The concrete method related to the story in the system is explained in Section 4.4. Figure 5 shows the time sequence of the processing of the interaction manager.

(4) *Each type of handler*: The handlers exist below the scene manager and the interaction manager, and have functions to control the various inputs and outputs. Concretely speaking, we have prepared the following handlers.

a) Speech recognition handler: This handler controls the speech recognition function. Speech recognition is carried out by software [20] developed at ATR. It is based on a Hidden Markov Model (HMM) and has a function for speaker-independent continuous speech recognition.
b) Emotions recognition handler: This handler controls the recognition of emotions included in voices. Emotions recognition is carried out by software [14] developed by the authors when researching computer characters reacting to emotions. A neural net is used as the basic algorithm for the emotions recognition. These emotions include eight types: happiness, anger, surprise, sadness, disgust, teasing, fear, and normality. The handler performs the recognition of emotions by a combination of eight types of neural nets trained for each of the emotions by using voice data from a number of speakers uttering 100 phoneme-balanced words.
c) Gesture recognition handler: This handler controls the recognition of body motions. The recognition of gestures is carried out by software called "Pfinder" [23] developed at MIT. Pfinder extracts silhouettes of characters from images taken from cameras, and has a function enabling it to execute processing to extract the head, both arms, and both legs (i.e., five characteristic points) in real time on an SGI Indy.
d) Image handler: This handler controls the output of images, such as the background and character animations.
e) Sound handler: This handler controls the output of sounds and voices, such as the background sound, effect sound, and lines of the characters.

4.3. Hardware Construction

Figure 6 shows the hardware construction. The hardware construction consists of an image output sub-system, voice and emotions recognition sub-system, gesture recognition sub-system, and sound output sub-system.

(1) *Image output sub-system*: This sub-system uses a high-speed workstation (Onyx Infinite Reality) for computer graphics creation as a workstation for image output. This workstation accommodates the script manager, scene manager, interaction

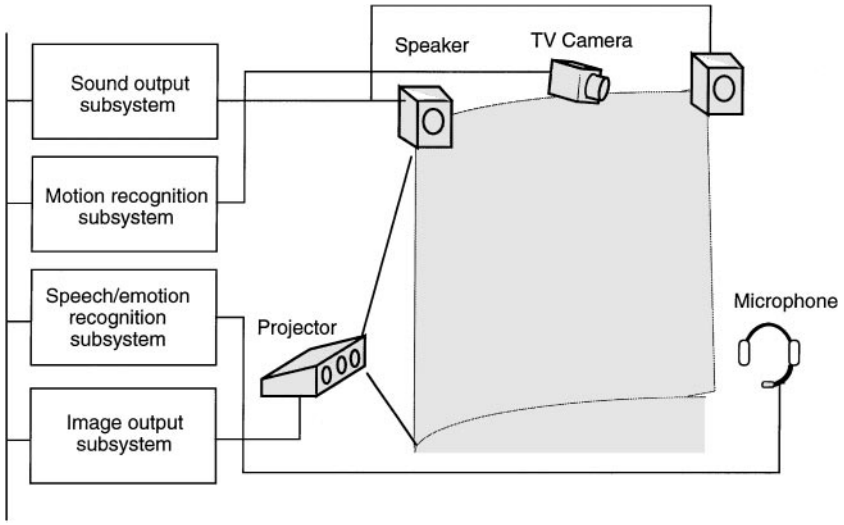


Figure 6. Hardware configuration of the Interactive Movie system.

manager, and each of the software packages of the image handler. The images of characters are stored beforehand as computer graphics animation data, and computer graphics is created in real time. Even computer graphics images of backgrounds are stored as digital data, and background images are created in real time. A part of each background image uses actual images, and they are stored in an externally connected LD. For all of these (computer graphics of many characters, computer graphics of backgrounds, and actual images of backgrounds), overlap processing takes place on a video board of ONYX.

For image creation rich to the feeling of presence, all computer images are displayed in 3D. In addition, for the participants to become immersed in the world of interactive movies surrounded by images, an arch-type screen is adopted. Beforehand, two types of image data (i.e., for the left eye and for the right eye) are created on a workstation, and while the data is being mixed through the use of a 3D observation controller, the information is projected onto the arch-type screen through two projectors (Fig. 7).

(2) *Speech and emotions recognition sub-system*: Speech and emotions recognition are carried out on one workstation (SUN SS20), on which the speech recognition handler and emotions recognition handler are also implemented. Voice inputs from a microphone are A/D converted from the internal SUN speech board, and speech recognition and emotions recognition are carried out through the installed speech recognition software and emotions recognition software.

(3) *Gesture recognition sub-system*: Gesture recognition is carried out on an SGI Indy, on which the gesture recognition handler is also implemented.

(4) *Sound output sub-system*: The sound output subsystem is constructed through a number of personal computers. The sounds that have to be output at the same time, are the background sound, the sound effects, and the lines of the characters. The effective sound and the lines of the characters are stored as digital data, and D/A is carried out when necessary. To support the simultaneous output of these sounds, the simultaneous D/A of three channels has been made possible. In addition, the background sound is recorded beforehand on an external CD, and output control is performed from a PC. The sounds output through these multiple channels are mixed and then output using a computer-controllable mixer (Yamaha O2R).

4.4. Construction Example of an Interactive Story

We constructed a concrete interactive story on the above considerations and system construction. As our concrete story, we adopted a representative old tale of Japan, i.e., Urashima Taro. The reasons are as follows.

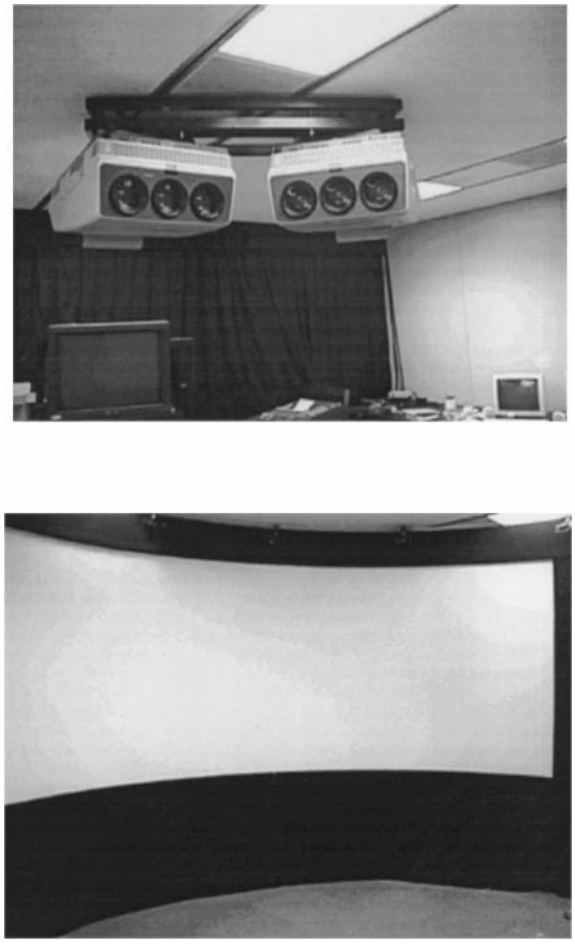


Figure 7. The projectors and the arch screen.

(1) Because it is a story understood by everyone, it is easy to become immersed in the story.
(2) Because it is an easy story that nevertheless holds a profound meaning, it is easy to arrange.
(3) It includes content that easily draws the interests of people (with gorgeousness image-wise, e.g., Dragon Palace).

Moreover, for the purpose of emphasizing the entertainment aspect in a cyberspace, we implemented an arrangement allowing some changes to be made: the turtle to a rabbit, Princess Oto-hime to a character called "MUSE" (goddess of beauty), and Dragon Palace to MUSE's Palace. Figure 8 shows the flowchart for the storytelling.

As Urashima Taro is a simple story, there are only four turning points where the story line changes depending on interaction, as given below.

(Turning Point 1) Whether or not to save the rabbit being tormented by some children.
(Turning Point 2) Whether or not to enjoy the entertainment provided at MUSE's Palace.
(Turning Point 3) Whether to remain at MUSE's Palace or return home.
(Turning Point 4) Whether or not to open the treasure box.

For each of these turning points, we investigated by primary testing with what modality was it easy for the participants to interact, and which movements and voices were easy to produce, and we classified the ideal interaction to each turning point.

For example, Turning Point 1 is a scene in which three young trouble-makers pick on the rabbit; participants who feel like stopping the three can easily produce such voices as "Stop it!", "Isn't that cruel?", or some other expression. Here, we selected several easy-to-produce sentences for both cases of yes and no. In addition, we made it possible to employ the emotions recognition results to handle cases in which sentences not in the dictionary are uttered. Decisions are made depending on whether an emotion recognition result corresponds to a positive or negative emotion. Figure 9 shows the scene for turning point 1.

Turning Point 4 is a scene of whether or not to open the treasure box received from MUSE. When a participant extends his/her hand upon deciding to open the box, this is determined as "yes" (to open the box); when the participant does nothing, this is determined as "no" (to not open the box). By displaying the treasure box in front of the participant in 3D, we induce the feeling of wanting to touch the box by extending the hand. Figure 10 shows an aspect of interaction for turning point 4.

Each participant wears liquid crystal shutter glasses for 3D observation and stands in front of the screen holding a microphone. While viewing the story displayed with images and sounds, he/she experiences interaction with characters in the movie, and experiences participating in the storytelling. As mentioned above, the participant interacts by the use of voices and gestures at the turning points in the storytelling. Of course, it is possible for the participant to make selections, e.g., whether to remain at the palace or return home, whether to open the treasure box or not, etc. Therefore, because of this, the content of the story will differ from the original.

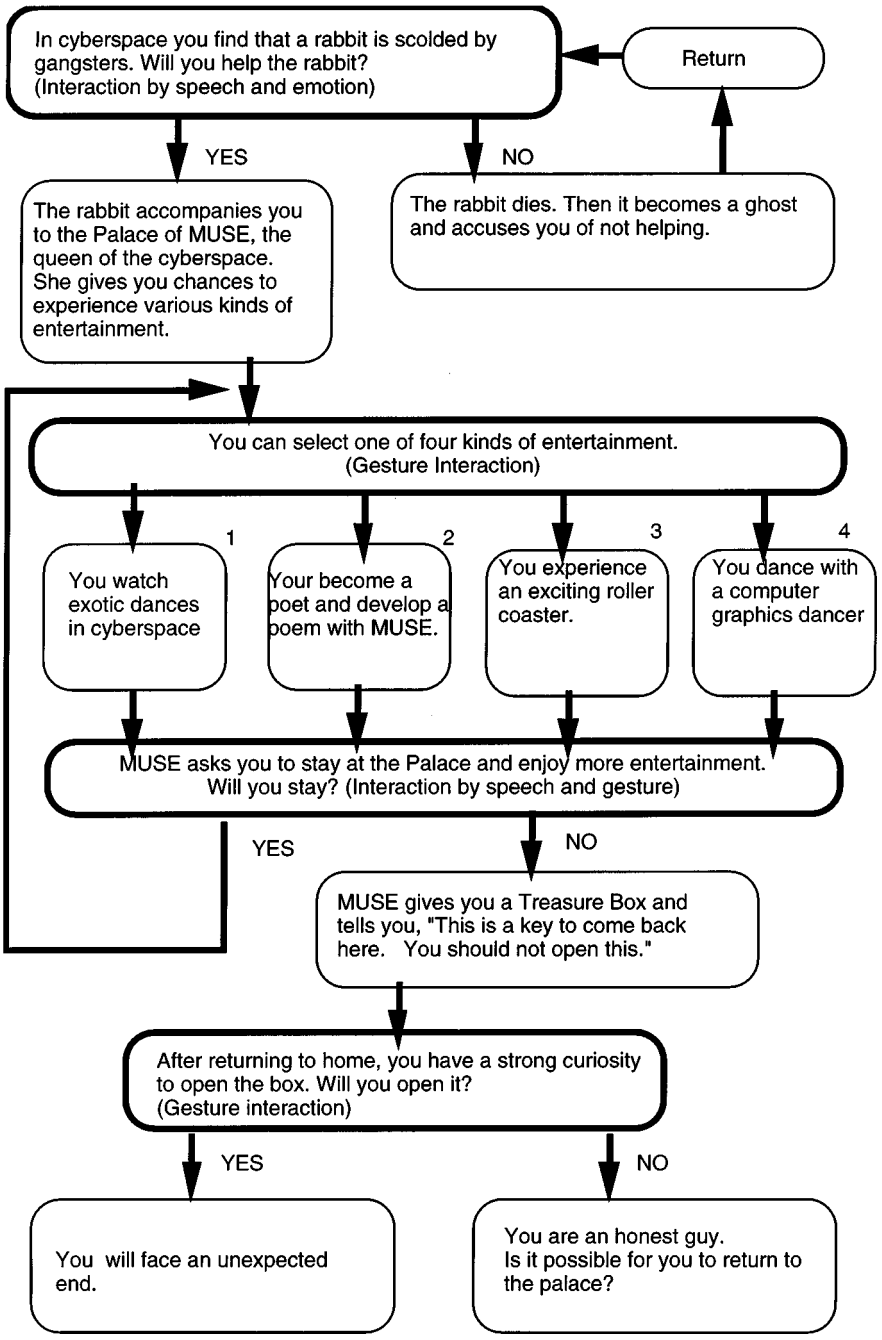


Figure 8. Interactive story example.

4.5. Evaluation and Problems

We tested the first prototype system with approximately 50 people during a half-year period following the completion of system development. Based on their comments, we evaluated the system and identified areas for further research, as summarized below.

(1) *Frequency of interaction*: Interaction was generally limited to the "change" points in the story, so the story progressed linearly along the

Figure 9. An example of scenes in the Interactive Movie system.

Figure 10. An example of interactions between the Interactive Movie system and a participant.

predetermined course, like a movie except at these change points. Using fixed story elements, created in the same way as for a conventional movie, was found to be disadvantageous, in that the participants seemed to end up being spectators and found it difficult to participate interactively at those points where interaction was clearly required. The limited opportunities for interaction in turn created other drawbacks for the participants, such as having little to go on to distinguish their experience in watching a movie, and having a very limited sense of involvement.

(2) *Number of participants*: The basic concept of this first system is a story with just one player acting the role of hero. This system lacks a multi-user functionality needed for the story to take place in a cyberspace, since cyberspaces exist over networks and require a story to develop from not just one player but from several players interacting at the same time.

5. Description of the Second Prototype System

5.1. Improvements

Based on the evaluation results, the following points were used to improve the first system as described below.

(1) *Introduction of spontaneous interaction*: To increase the frequency of interaction between the participants and the system, we devised a way for the participants to interact with movie world residents at any point in time. Basically, these impromptu interactions occur between the participants and characters and generally do not affect story development. On the other hand, there are interactions at times that do affect story development. Such interactions occur at branch points in the story, and consequently, they tend to determine the future development of the story. Therefore, to achieve spontaneous interaction, the key point is how to handle these two types of interactions.

(2) *System for multiple players*: Our initial effort to develop a system (i.e., second prototype system) for multiple players has made it possible for two players to participate in a narrative world in the development of a story. The ultimate goal is to create a multi-player system operating across a network. The first step in this study was the development of a prototype multi-player system consisting of two systems connected by a LAN.

(3) *Other improvements*:

Emotions recognition: To achieve spontaneous interaction, the usage of the emotions recognition capability was extended. When the participants utter spontaneous utterances, characters react by using their own utterances and animations according to the emotion recognition result.

Motion capture: We introduced a motion capture system based on magnetic sensors. There are two major reasons for implementing such a system. One is to show avatars as the alter egos of the participants on the screen, thereby giving the participants the feeling that they are in fact active participants in the system. The other is to improve the recognition of gestures. As the first system's gesture recognition, based on images obtained by a camera, was ineffective due to low light, we adopted the use of motion capture data for gesture recognition.

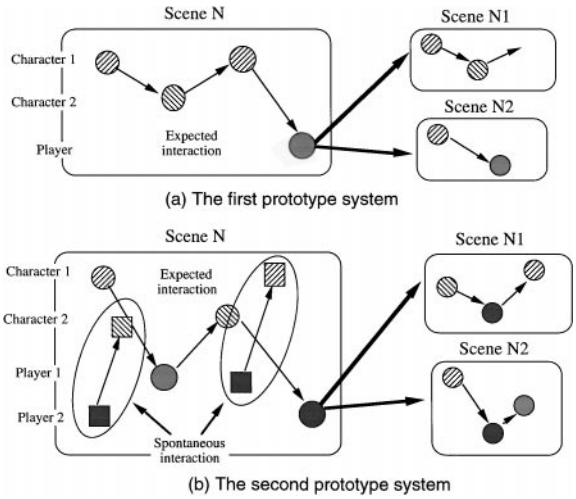


Figure 11. Comparison of interactions between the first system and the second system.

5.2. Software System Structure

(1) *System structure concept*: While the first system stresses story development, the second system strives to achieve a good balance between story development and impromptu interaction by incorporating the concept of spontaneous interaction.

Figure 11 schematically illustrates how interaction proceeds for both the first system and the second system. In the first system, the order of all interactions of a participant and the behaviors of characters are predetermined. In addition, at most one interaction is included in one scene. Under such conditions, it can be understood that the system control mechanism is rather simple. Figures 4 and 5 illustrate how the scene manager and the interaction manager work in the first system.

The difference between the first system and the second system, as illustrated in Fig. 11, is that in addition to the predetermined sequence of interactions between the participants and characters, unpredictable interactions, in other words, spontaneous interactions, occur. Therefore, the second system is required to distinguish predicted and unpredicted interaction results and to handle the behaviors of the characters corresponding to these two kinds of inputs. Furthermore, to add naturalness to the reactions of the characters, some

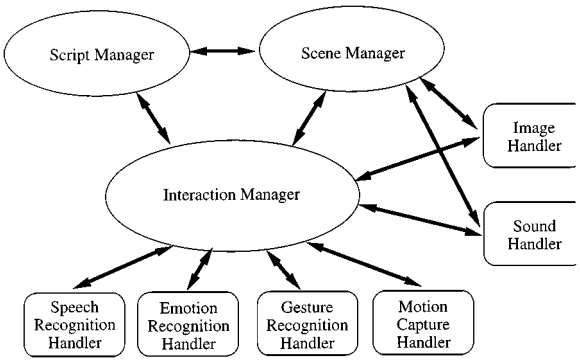


Figure 12. Software configuration of the second system.

fluctuations are added for the response of each character. These requirements have led to the building of a distributed control system instead of a top-down system structure. Figure 12 illustrates the structure of the software used in the second system.

(2) *Script manager*: The role of the script manager is to control the transitions between scenes, just as it did with the first system. An interactive story consists of various kinds of scenes and transitions between scenes. The function of the script manager is to control these scene transitions based on an infinite automaton (Fig. 2). The transition from a single scene to one of several possible subsequent scenes is determined based on the interaction result sent from the scene manager.

(3) *Interaction manager*: The interaction manager is the most critical component for achieving spontaneous interaction. The interaction manager should therefore have functions for distinguishing predicted and unpredicted (spontaneous) interactions and generating character reactions for spontaneous interactions. The details of these functions are described below.

a) Distinguishing predicted and unpredicted interactions: It is a difficult issue for the system to distinguish whether the speech/gesture of a participant is a predicted input or unpredicted input by only using speech recognition technologies. To solve this problem, we have adopted a method in which both speech recognition and emotion recognition work simultaneously. When a reasonable speech recognition result is obtained, the input is judged as a predicted interaction and the recognition result is utilized for generating predetermined behaviors of the characters. When the speech recognition function fails to output a reasonable recognition result, in contrast, this is judged as a spontaneous interaction, and then, the emotion recognition result is utilized to generate spontaneous reactions of the characters (described below). For gesture recognition, because the usage of predetermined gesture interactions is restricted, it is rather easy to distinguish between two kinds of gesture inputs depending on the scene.

b) Generating character reactions for spontaneous inputs: The basis of spontaneous interaction is a structure that allots each character an emotional state, and the interaction input from the participant combines his/her interaction with the characters to determine the emotional state (as well as the response to that emotional state) of each character. Some leeway is given to how a response is expressed depending on a character's personality and circumstances. To achieve this, we developed an algorithm based on the concept of the "action selection network" [22] which sends and receives activation levels among multiple nodes. How the interaction manager works for spontaneous inputs is outlined below.

The state and intensity of a participant's ($i = 1, 2$) emotion at time T is defined as

$$\mathrm{Ep}(i, T), \mathrm{sp}(i, T) \text{ where } \mathrm{sp}(i, T) = 0 \text{ or } 1$$

(0 indicates no input and 1 indicates an input).

Similarly, the state and intensity of a character's ($j = 1, 2 \ldots$) emotion at time T is defined as

$$\mathrm{Eo}(j, T), \mathrm{so}(j, T).$$

The emotional state of character j is determined as a function of the emotional states of the participants and the character:

$$\mathrm{Eo}(j, T+1) = \mathrm{f1}\,(\mathrm{Ep}(1, T), \mathrm{Ep}(2, T), \mathrm{Eo}(j, T)).$$

When the emotion of participant i is recognized, the activation level sent to character j is determined as a function of $\mathrm{sp}(i, T)$ and j:

$$\mathrm{sp}(i, j, T) = \mathrm{f2}\,(\mathrm{sp}(i, T), j)$$

where $\mathrm{sp}(i, j, T)$ is the activation level sent to character j. The activation level for character j

is the total of all activation levels received by the character:

$$\mathrm{so}(j, T+1) = \mathrm{sp}(1, j, T) + \mathrm{sp}(2, j, T) + \mathrm{so}(j, T).$$

A character that exceeds the activation threshold performs action $\mathrm{Ao}(j, T)$ based on an emotional state. More specifically, this action involves the character's movement and speech as a reaction to the emotional state of the participant. At the same time, activation levels $\mathrm{so}(j, k, T)$ $(k = 1, 2 \ldots)$ are sent from character j to other characters and then its activation level so (j, T) is set to zero:

if $\mathrm{so}(j, T) > T\mathrm{Hi}$
then $\mathrm{Ao}(j, T) = \mathrm{f3}\ (\mathrm{Eo}(j, T), j)$,
$\mathrm{so}(j, k, T) = \mathrm{f4}\ (\mathrm{so}(j, T), k)$
$\mathrm{so}(k, T+1) = \mathrm{sum}\ \{\mathrm{so}(j, k, T)\} + \mathrm{so}(k, T)$.
$\mathrm{so}(j, T+1) = 0$

This mechanism creates interaction between characters and enables more diverse interaction than simple interaction involving a one-to-one correspondence between emotion recognition results and character reactions.

c) Other issues:

Time control: A difficult issue in handling spontaneous interaction is that once we permit this, controlling the time schedule for the sequence of predetermined interactions becomes difficult, because the scenario as a whole is controlled by the scene data handled by the scene manager. As one solution to this problem, we introduced the concept of the "relative time counter". Here, while characters are showing reactions corresponding to spontaneous interaction, the timer stops counting. This means that as long as the participants continue to enjoy the spontaneous interaction, the story stops proceeding. By introducing this mechanism, the system can go to any point between a fully spontaneous interaction system and a fully narratively controlled system.

Reaction collision: There are cases where a spontaneous input comes in while a character is reacting to an expected interaction of a participant. In this case, a collision occurs between the expected reaction and the reaction to the unexpected reaction.

We employ two modes in this case: real-time reaction and delayed reaction. In the case of real-time reaction, the character puts the reaction on hold and instead shows the spontaneous reaction. In the case of delayed reaction, the character continues its behavior and, after finishing it, starts its reaction for the spontaneous input.

(4) *Scene manager*: The scene manager controls the generation of scenes as well as the progress of the story in a scene. To control the ongoing progress of each scene, we define the scene data, which controls all of the events for each scene. A brief description of the scene data construction is as follows.

The scene data consists of the following kinds of commands and parameters.

***MACRO COMMAND
COMMAND
Parameter

Macro commands control the description of a scene. The macro commands we have prepared are given in the following.

***SCENE: command to define the scene number
***TIMER: command to define the maximum duration time for each scene
***CAMERA: command to define the camera position for each scene

Commands define and describe the details of events in each scene. The main commands for the present system are given.

BGIMG: command to define the background image
SOUND: command to define the background music and sound effects
OBJECT: command to define the computer graphics corresponding to various kinds of objects and characters
MOTION: command to define the motion capture mechanism
INTERACTION_S: command to define the interaction based on speech recognition
INTERACTION_G: command to define the interaction based on gesture recognition

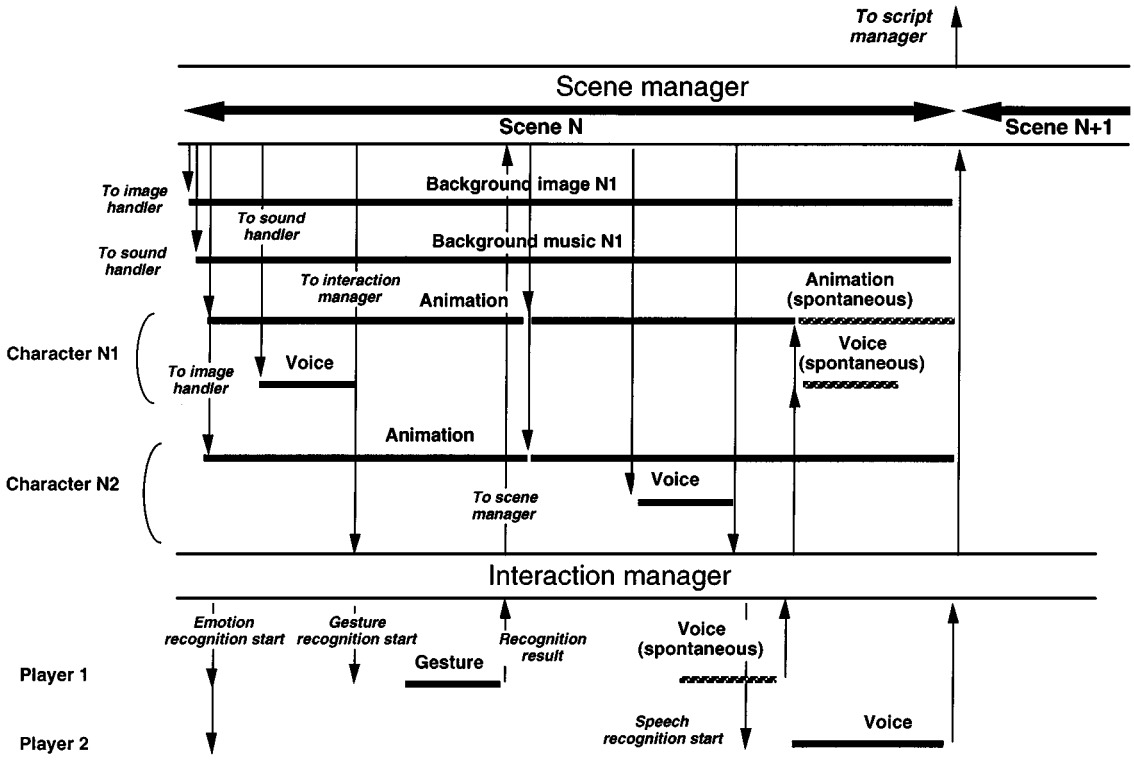


Figure 13. Time sequence of scene/interaction manager processing (second prototype system).

INTERACTION_E: command to define the interaction based on emotion recognition

Parameters define various kinds of conditions for each command. An example of a scene data description for a spontaneous interaction is as follows

INTERACTION_E			
Character_N	Start_time	End_time	
Participant_M1	Wait	File_CG(M1)	File_speech(M1)
Participant_M2	Imm.	File_CG(M2)	File_Speech(M2)

This means that during the time between the Start_time and End_time, the character indicated by N can accept spontaneous inputs from participants M1 and M2. When reactions are activated based on the mechanism mentioned in (3), the reaction of participant M1 is expressed by File_CG(M1) for animation and File_speech(M1) for speech. In addition, the type of reaction is defined by Wait or Imm, for the delayed or immediate reactions described in (3), respectively.

Figure 13 illustrates how the interaction manager and scene manager work, and as a result, how the interaction between the participants and characters proceeds.

5.3. Hardware System Structure

Figure 14 shows the second system's hardware structure, composed of image output, voice and emotion recognition, gesture recognition, and sound output subsystems.

(1) *Image output sub-system*: Two workstations (Onyx Infinite Reality and Indigo 2 Impact) capable of generating computer graphics at high speed are used to output images. The Onyx workstation is used to run the script manager, scene manager, interaction manager, and the entire image output software.

Character images are pre-stored on the workstations in the form of computer graphics animation data in order for the computer graphics to be generated in real time. Background computer graphics images are also stored as digital data enabling the

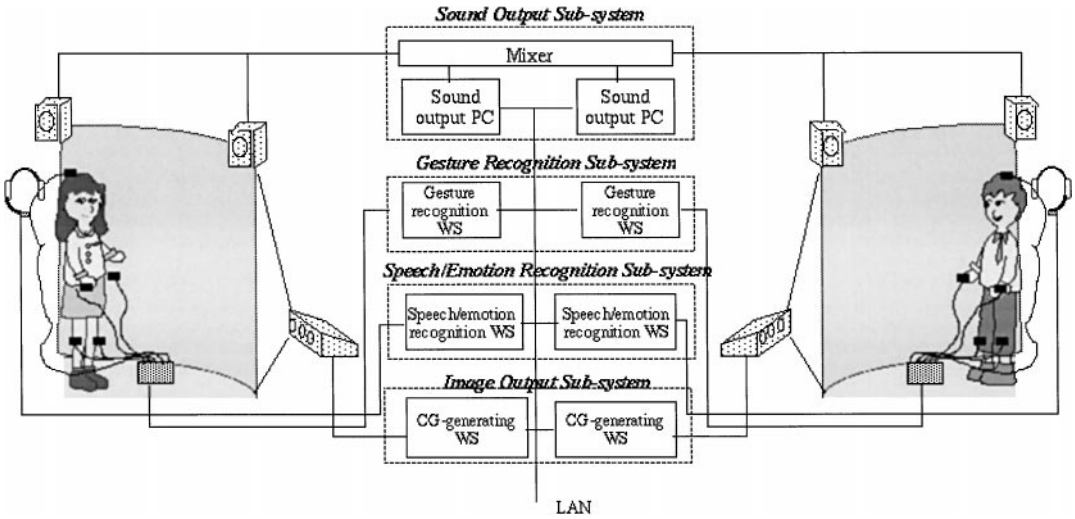


Figure 14. Hardware configuration of the second system.

background images to be generated in real time. Some background images are photographic images of real scenery stored on an external laser disc. The multiple character computer graphics, background computer graphics, and background photographic images are processed simultaneously through video boards on both the Onyx and Indigo 2 workstations.

The computer graphics images are displayed in 3-D to give a more realistic sensation, and a curved screen is used to envelop the participant with the images and immerse him or her in the interactive movie world. The image data for the left eye and right eye, previously created on the workstations, are integrated by stereoscopic vision control and projected onto the curved screen by two projectors. On the Indigo 2 end, however, the images are output on an ordinary large-screen display without stereoscopic vision because of processing speed limitations.

(2) *Voice and emotion recognition sub-system*: Voice and emotions are recognized with two other workstations (Sun SS20s) which also run the voice and emotion recognition handlers. Voice inputs via a microphone are converted from analog to digital by soundboards built into the Sun workstations, and the recognition software on the workstations is used to recognize voices and emotions. For the recognition of meanings, a speaker-independent speech recognition algorithm based on an HMM is adopted [24]. Emotions recognition is achieved by using a neural-network-based algorithm [14]. Each workstation processes the voice inputs from one participant.

(3) *Gesture recognition sub-system*: Gestures are recognized with two SGI Indy workstations that run the gesture recognition handler. Each workstation takes the output from the magnetic sensors attached to a participant and uses that data to control the avatar of the participant and to recognize gestures.

(4) *Sound output sub-system*: The sound output subsystem uses several personal computers because of the need to output background music, sound effects, and speech for each character simultaneously. The sound effects and character speech are stored as digital data that are converted from digital to analog as needed, and multiple personal computers are used to enable simultaneous digital to analog conversion of multiple channels in order to output these sounds simultaneously. The background music is stored on an external compact disc whose output is also controlled by a personal computer. The multiple-channel sound outputs are mixed and output with a mixer (Yamaha 02R) that can be controlled by a computer.

5.4. Example of Interactive Story Production

(1) *An interactive story*: We produced an interactive story based on the previously described Second System. We selected “Romeo and Juliet” by Shakespeare as the base story for the following reasons.

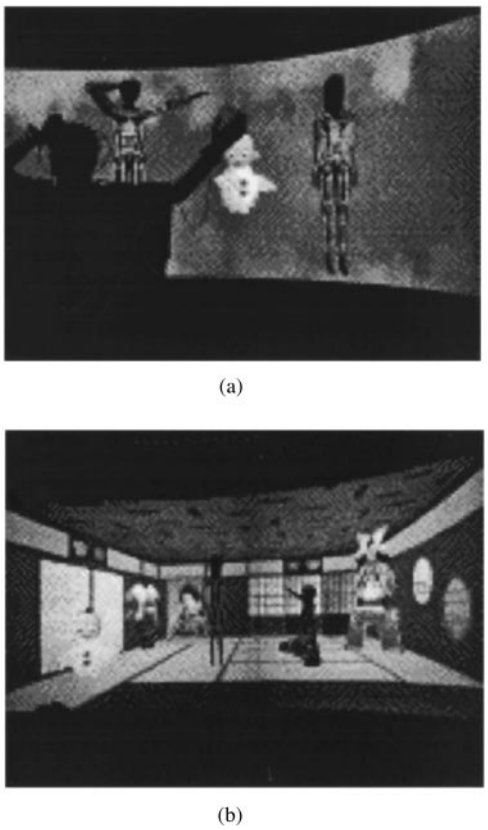
a) There are two main characters in the story and, because of this, the story supplies a good example of multi-person participation.
b) "Romeo and Juliet" is a very well known story, and people have a strong desire to act out the role of the hero or heroine. Therefore, it is expected that people can easily get involved in the movie world and experience the story.

The main plot of the story is as follows. After their tragic suicide the lovers' souls are sent to Hades, where they find that they have totally lost their memory. Then, they start their journey to rediscover who they are and what their relationship was. With various kinds of experiences and with the help and guidance of characters in Hades, they gradually find themselves again and finally go back to the real world.

(2) *Interaction*: There are two participants; one plays the role of Romeo and the other of Juliet. The two sub-systems are located in two separate rooms and connected by a LAN. Each participant stands in front of the screen of his/her respective system wearing specially designed clothes to which magnetic sensors and microphones are attached. In the case of Romeo, the participant wears a 3-D LCD-shutter glass and can enjoy 3-D scenes. Their avatars are on the screen and move according to their actions. They can also communicate by voice. Basically, the system controls the progress of the story with character animations and character dialogues. Depending on the voice and gesture reactions of the participants, the story moves on. Furthermore, as was described before, interaction is possible at any time. When the participants speak, the characters react according to the emotion recognition results. Consequently, depending on the frequency of the participants' interaction, this system can go anywhere from story-dominant operation to impromptu interaction-dominant operation. Figure 15 illustrates typical interactions between the participants and the system. Also Fig. 16 shows examples of the possible dialogue sequences and scene changes.

5.5. *Considerations*

With the second system, researchers at the laboratories and visitors to the laboratories (about 50 people) were able to gain experience over a period of about six months since the development of the system. It is necessary to do formal evaluation experiments in different ways, but from the impressions of these people who gained experience we could collect several suggestions concerning the evaluation of the system and how we will progress with the system in the future.



(a)

(b)

Figure 15. Examples of interaction between a participant and the system. (a) "Romeo" controls his avatar, (b) "Romeo" tries to touch an object in a Japanese gift shop.

(1) *Comparison with video games*: We could discriminate the second system from video games as follows. The interactive movie system is superior to video games because of the power through a large screen and 3D observation. Moreover, its interaction function, which involves voices and movements (not button inputs), lead most people easily into deep engagement into story development. On the other side, there were also opinions stating that a longer story would give people deeper emotions and immersion.
(2) *Comparison with movies*: Many participants felt that it was a new experience to be the main figure in storytelling, instead of the conventional way of

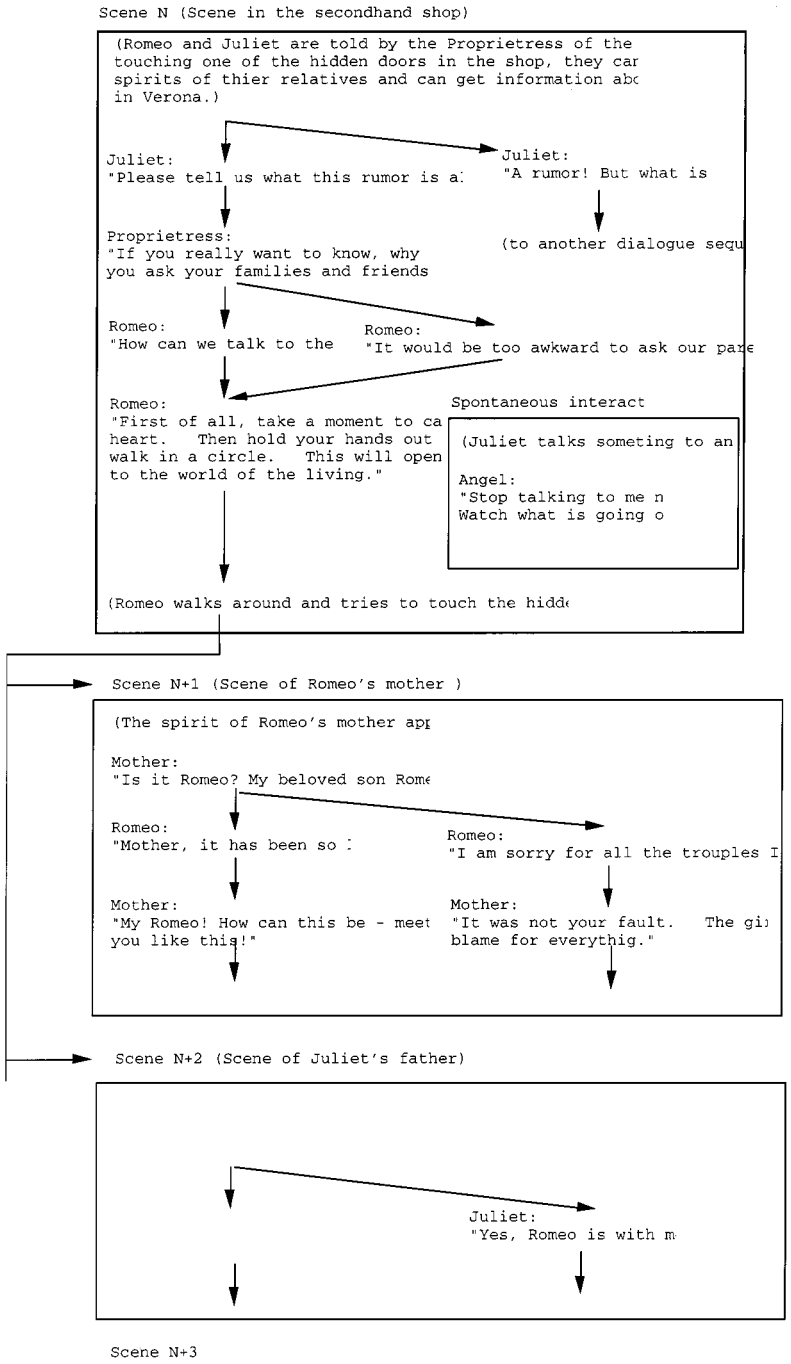


Figure 16. Examples of dialogue sequence between participants and the system.

appreciating movies passively. A small part of the participants remained as passive type participants and did not actively interact because they were not used to this new type of experience.

These results showed that further improvements are required concerning the interaction and the construction of the story, but they also showed the possibility of interactive movies as a new type of media.

6. Conclusion

We have explained Interactive Movies, which can be thought of as a new type of media combining various types of media such as communications, movies, video games, etc. Interactive Movies are a new type of media that construct cyberspaces rich to the feeling of presence through computer graphics and 3D observation, and participants can enter into these cyberspaces to experience the storytelling while interacting with other characters.

In this paper, we have proposed our concept of Interactive Movies, and along with this, considered their relationship with past media as well as the conditions for achieving them. Then, we explained our first prototype system. Based on an evaluation of this system, we identified several problems in the system that needed to be improved. One was the lack of frequent interactions and the other was single-person participation. To overcome these deficiencies, we are developing a second system. We explained two significant improvements incorporated into the second system: spontaneous interaction and two-person participation through a network. We described the software and hardware configurations of the second system while emphasizing these improvements. Finally, we illustrated the operation of our second system with the example of our interactive production of "Romeo and Juliet".

References

1. H. Rheingold, *The Virtual Community: Homesteading on the Electronic Frontier*, Reading, Mass: Addison-Wesley, 1993.
2. O. Jacob, "Computer Graphics Story—A Personal Overview of Computer Animation in the Movies," *ACM Computer Graphics*, vol. 31, no. 1, 1997, pp. 26–28.
3. G. Walters et al., "The Making of Toy Story," Course Notes of SIGGRAPH 96, Aug. 1996.
4. J.H. Murray, *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*, New York: Simon & Schuster, 1997.
5. S. Turkle, *Life on the Screen: Identity in the Age of the Internet*, New York: Simon & Schuster, 1995.
6. R. Nakatsu and N. Tosa, "Toward the Realization of Interactive Movies—Inter Communication Theater: Concept and System," in *Proceedings of the International Conference on Multimedia Computing and Systems'97*, June 1997, pp. 71–77.
7. R. Nakatsu and N. Tosa, "Real-time Spontaneous Interaction System with Narratives," in *Proceedings of 1998 IEEE Second Workshop on Multimedia Signal Processing*, 1998, pp. 247–252.
8. C.W. Ceram, *Eine Archaologie des Kinos*, Hamburg: Rowohlt, Verlag, 1965.
9. H. Noma et al., "Multi-Point Virtual Space Teleconference System," *IEICE Trans. Commun.*, vol. E78-B, no. 7, July 1996.
10. G. Davenport, "Smarter Tools for Storytelling: Are They Just Around the Corner?" *IEEE Multimedia*, vol. 4, no. 1, 1996, pp. 10–14.
11. B. Laurel, *Computers as Theater*, Reading, MA: Addison-Wesley, 1993.
12. J. Weizenbaum, "ELIZA: A Computer Program for the Study of Natural Language Communication between Man and Machines," *Communications of the ACM*, no. 9, 1996, pp. 36–45.
13. P. Maes et al., "The ALIVE System: Full-Body Interaction with Autonomous Agents," in *Proceedings of the Computer Animation '95 Conference*, 1995.
14. N. Tosa and R. Nakatsu, "Life-like Communication Agent—Emotion Sensing Character 'MIC' and Feeling Session Character 'MUSE'," in *Proceedings of the International Conference on Multi-media Computing and Systems*, June 1996, pp. 12–19.
15. J. Bates et al., "An Architecture for Action, Emotion, and Social Behavior," in *Proceedings of the Fourth European Workshop on Modeling Autonomous Agents in a Multi-Agent World*, July 1992.
16. K. Perlin, "Real Time Responsive Animation with Personality," *IEEE Trans. on Visualization and Computer Graphics*, vol. 1, no. 1, 1995, pp. 5–15.
17. *Proceedings of Life-like Computer Characters'96*, Oct. 1996.
18. C. Sommerer and L. Mignonneau (Eds.), *Art@Science*, Wien: Springer-Verlag, 1998.
19. C. Pinhanez and A. Bobick, "It/I: A Theater Play Featuring an Autonomous Computer Graphics Character," M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 455, 1998.
20. *Commun. ACM Special Issue on Intelligent Agents*, vol. 37, no. 7, 1994.
21. A. Bruderlin and T. Calvert, "Knowledge-driven, Interactive Animation of Human Running," *Graphics Interface'96*, May 1996, pp. 213–221.
22. S. Weitz, *Nonverbal Communication*, New York: Oxford Univ. Press, 1974.
23. C.R. Wren et al., "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, July 1997, pp. 780–785.
24. T. Shimizu et al., "Spontaneous Dialogue Speech Recognition Using Cross-Word Context Constrained Word Graph," in *Proceedings of ICASSP'96*, Vol. 1, April 1996, pp. 145–148.
25. P. Maes, "How to do the Right Thing," *Connection Science*, vol. 1, no. 3, 1989, pp. 291–323.

Ryohei Nakatsu received the B.S., M.S. and Ph.D. degrees in electronic engineering from Kyoto University in 1969, 1971 and 1982

respectively. After joining NTT in 1971, he mainly worked on speech recognition technology. Since 1994, he has been with ATR (Advanced Telecommunications Research Institute) and currently is the president of ATR Media Integration & Communications Research Laboratories. His research interests includes emotion extraction from speech and facial images, emotion recognition, nonverbal communications, and integration of multi modalities in communications. He is a member of the IEEE, the Institute of Electronics, Information and Communication Engineers Japan (IEICE-J), the Acoustical Society of Japan, Information Processing Society of Japan, and Japanese Society for Artificial Intelligence.
nakatsu@mic.atr.co.jp

Naoko Tosa is a Media Artist & Researcher in the ATR Media Integration & Communications Research Laboratories. She is also a lecturer in the Dept. of Imaging Arts and Sciences, Musashino Art University. Her major research area is Art and Technology where she is working on the creation of film & video, computer graphics animations, and interactive arts. Her recent work includes the Neuro-Baby project, an autonomous computer agent with automatic facial expression and behavior synthesis that can respond to human voice by recognizing emotions and feelings. Her work was exhibited at Museum of Modern Art (New York), Metropolitan Art Museum, SIGGRAPH, Ars ELECTRONICA, Long Beach Museum, and other locations worldwide. Also, her works are collected at The Japan Foundation, American Film Association, Japan Film Culture Center, Nagoya Prefectural Modern Art Museum Japan, and other institutions in Japan.
tosa@mic.atr.co.jp

Takeshi Ochi received the B.S. degree from Kinki University in 1989. In 1989 he joined CSK Inc. From 1996 he is working as a software engineer at ATR Media Integration & Communications Research Laboratories. His main area is computer graphics and image processing.
ochi@mic.atr.co.jp