Real-time Human Proxy : An Avatar-based Interaction System

Arita, Daisaku
Department of Intelligent Systems, Kyushu University

Yoshimatsu, Hisato
Department of Intelligent Systems, Kyushu University

Hayama, Daisuke
Department of Intelligent Systems, Kyushu University

Kunita, Masashi
Department of Intelligent Systems, Kyushu University

他

出版情報：Proceedings of International Conference on Multimedia and Expo '04. 3, pp.1835-1838, 2004-06
バージョン：
権利関係：

REAL-TIME HUMAN PROXY: AN AVATAR-BASED INTERACTION SYSTEM

Daisaku Arita, Hisato Yoshimatsu, Daisuke Hayama, Masashi Kunita, Rin-ichiro Taniguchi

Dept. of Intelligent Systems, Kyushu Univ.
6-1, Kasuga-koen, Kasuga, Fukuoka 816-8580 Japan
{arita, hisato, daizz, mkunita, rin}@limu.is.kyushu-u.ac.jp

ABSTRACT

In this paper, we propose a concept of real-time human proxy for avatar-based interaction systems, which virtualizes a human in the real world in real-time and which lets the virtualized human behave as if he/she was present at a distant place. For estimating RHP, we apply it to VEIDL, which is a virtual classroom system. The experimental results shows us that RHP is useful for avatar-based interaction.

1. INTRODUCTION

A lot of network-based human interaction systems have been developed. These systems handle video and sound streams, which are captured by a camera and a microphone at each site, transmitted via a network, and presented by a video display and a speaker. The disadvantages of such audio-visual interaction systems are summarized as follows.

Inconsistency of positional relations
: Since all visual information is presented on a 2-D display, positional relations among participants are not consistent. This means that, for example, each participant can not understand where other participants look at and point to.

Limitation of the number of participants
: Since video images of all participants are arranged on a 2-D display, the number of participants is limited by the size and the resolution of a display.

Privacy
: Audio-visual interaction systems convey participants' information which participants sometimes do not want to convey, such as their faces, their clothes and their rooms without concealment.

To solve the problems of video-based interaction systems, there are several researches on virtual environments for human interaction[1, 2]. In these researches, a 3-D virtual space is reconstructed, in which each participant is represented as an avatar generated by computer graphics. Through a reconstructed virtual space, each participant virtually see and hear other participants' activities from the position where his/her avatar is represented. Since the 3-D virtual space is reconstructed, positional relations between participants can be consistent. This means that each participant can understand where other participants look at and point to, see where he/she wants to see, understand where a sound comes from, and move in the virtual space. In addition, since a display presents not video images of participants but a single virtual space, there is, in principle, no limitation of the number of participants caused by visibility[1].

However, it is difficult for avatar-based interaction systems to acquire and present all of participants' information, especially nonverbal information. For example, motion capture systems cannot acquire all motion information such as the angles of fingers. On the other hand, it is not necessary for an avatar to act just the same as participant's motions. In this paper, we propose Real-time Human Proxy, a new concept for avatar-based interaction, which makes it easy to acquire and present participants' information.

2. REAL-TIME HUMAN PROXY

For natural avatar-based interaction, we have introduced a concept of Real-time Human Proxy(RHP), which virtualizes a human in the real world in real-time and which lets the virtualized human behave as if he/she was present at a distant place. RHP acquires verbal and nonverbal information, transmits acquired information of human activity, and presents transmitted information in real-time.

In this paper, we focus on acquisition and presentation of nonverbal information.

[1]Needless to say, there is a limitation caused by computational power and network bandwidth

0-7803-8603-5/04/$20.00

2.1. Information Acquisition

To acquire nonverbal information, we use a real-time motion capture system which we are constructing[3, 4]. The system uses multiple cameras around a human. By using the system, we can get 3-D positions of a head, hands, elbows, knees, feet and a torso in real-time. However, it cannot acquire enough information to let an avatar behave similar to a human since the system cannot acquire all motion parameters for an avatar including articular angles of wrists and fingers, and twist angles of shoulders and hip joints. This means that the system must compensate for motion parameters with pre-defined knowledge, which requires the system to recognize participant's intention from limited motion parameters acquired by the motion capture system. For example, in case that a participant walks, recognizing the motion as walking, transmitting minimum information, and synthesizing motion parameters of legs such as angles of knees, ankles and toes generates more natural walking scenes because fine motion parameters such as angles of ankles and toes can not be acquired by the motion capture system. Of course, it is reasonable not only for ability of motion capture but also for amount of network data transmission

For these reasons, the system transmits information in a symbol data form if possible. For example, in case of walking, symbol "walking (p_x, p_y)" is transmitted, where p_x and p_y are the position of the participant. However, this rule should be changed according to the situation. If walking parameters are important for the interaction such as rehabilitation and dance training, walking parameters such as angles of knees and ankles should be transmitted in a raw data form.

The criterion for deciding whether information should be transmitted in a raw data form or in a symbol data form is intentionality of participants avatars, which depends on the situation. Then, we establish a rule in advance describing which information should be transmitted in a raw data form or in a symbol data form according to the situation of the interaction. When information should be transmitted in a symbol data form, the system recognizes and symbolizes the human motion of the information and transmits the result.

2.2. Information Presentation

It is impossible for the system to present a human avatar referring to only transmitted information since symbolic representation has no detailed motion data of the body parts but just a label of action. This means that the system has to generate motion parameters of the body parts to present a human avatar referring to pre-defined knowledge, which represents the correspondence between symbolic representation and detailed motion parameters of the body parts. The knowledge is represented in *motion model*.

In addition, the system transmits only information about pre-defined actions. This means that no information is transmitted when a participant does not make any pre-defined actions. If an avatar acts only when information is transmitted, the avatar often freezes. Of course, such avatar's behavior is not natural. To solve this problem, the system has to fill action intervals between pre-defined actions for time-filling. The time-filling actions must be actions that do not indicate participants' intentions in order not to influence on the interaction among participants. For example, such actions as "folding arms", "sticking hand into a pocket" and "crossing legs" can be usually used as time-filling actions. This kind of knowledge is represented in *behavior model*.

2.2.1. Behavior Model

Behavior model is referred for deciding which action an avatar makes next time. Behavior model of an avatar is described as a set of state transition graphs. Each graph corresponds to a body part of an avatar such as a left arm and a right leg. Each state in the graph corresponds to an action of the part such as "raising hand", "folding arms", "standing" and "walking". Transition probabilities are manually defined in order that the avatar acts naturally.

When transmitted information is received, the current state of a graph is forced to transits to the state corresponding to the transmitted information. This is realized by changing transition probabilities in order to guide the current state to the state corresponding to the received information. This is because this mechanism guarantees not to change actions discontinuously and makes avatar's action natural. For example, in case that symbol "finger pointing by a right hand" is received when an avatar sticks its right hand into its pocket, the avatar takes its right hand out of its pocket and then points by finger.

In addition, transition probabilities are changed according to the current states of other graphs. This is because an action with multiple parts of an avatar such as "folding arms" and "walking" requires that the current states of the parts are synchronized with one another. This is realized by a mechanism that transiting to such a state changes transition probabilities of related graphs in order to guide the current state of each related part to the corresponding state. When the current states of all related parts arrive at the corresponding states, an avatar synchronously starts the action with multiple parts.

2.2.2. Motion Model

Motion model is pre-defined knowledge for generating motion parameters of avatar body parts according to avatar's action decided by behavior model. Motion model has a sequence of motion parameters for each action. In case of transmitted information with additional parameters such as 3-D positions, the inverse kinematics theory is used for generating motion parameters depending on the additional parameters. The system lets an avatar act according to a sequence of motion parameters corresponding to an action decided by behavior model.

3. EXPERIMENTS ON VEIDL

We are developing a prototype of RHP, called VEIDL (Virtual Environment for Immersive Distributed Learning). VEIDL is a virtual classroom environment where avatars of geographically dispersed participants are teaching and learning together as shown in figure. 1. Verbal and nonverbal information uttered by participants is acquired by cameras and microphones and transmitted via network with one another. Each participant can see the scene of a classroom generated by computer graphics from the viewpoint of his/her avatar and hear the synthetic sound of the classroom at the point of his/her avatar. This means that each participant interacts with others through the virtual environment, or the virtual classroom. The advantage of dealing with a virtual classroom is that it is easy to decide which information should be transmitted since the objective of interaction in a classroom is clear. In this section, we will show the first step of RHP for VEIDL.

3.1. Information Acquisition and Transmission

For the first step, we have decided that four kinds of basic nonverbal information is necessary for presenting teacher's intentions; "walking", "body direction", "face direction" and "finger pointing". The system measures and recognizes these action based on output results of our motion capture system we have developed. However, the system cannot measure teacher's face direction for the present[2]. Then, a teacher wears markers on his/her head for measuring the face direction.

3.2. Information Presentation

We defined the behavior model of four human parts for this experiment as shown in Figure. 2. Since motion parameters in the motion model are manually constructed, avatar's motion is a little unnatural. It will be solved by using motion parameters acquired by a motion capture system with markers and a hand shape measuring system.

[2]We are researching for it now.

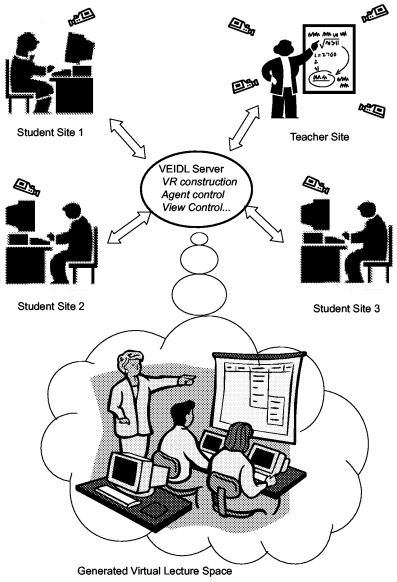


Fig. 1. Concept of VEIDL

3.3. Experimental results

Figure. 3 is original images and avatar images. These images show information acquisition performs well. Since motion model is primitive, avatar's motion is awkward. However, this problem will be solved by reinforcing motion model.

4. CONCLUSIONS

In this paper, we propose a concept of real-time human proxy for avatar-based interaction systems, which virtualizes a human in the real world in real-time and which lets the virtualized human behave as if he/she was present at a distant place. For estimating RHP, we apply it to VEIDL, which is a virtual classroom system. The experimental results shows us that RHP is useful for avatar-based interaction.

Future works are as follows.

Building VEIDL with RHP

Interaction in this experiment is only one way, a teacher to students. Then, it is impossible to estimate RHP as interaction mediator. For estimating RHP, we have to build VEIDL with two ways interaction.

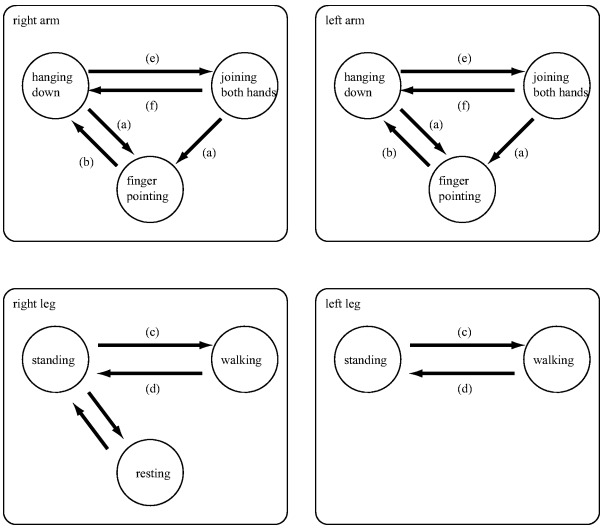


Fig. 2. Behavior model: Probability of transition (a) is changed from 0 to 1 when symbol "start finger pointing" is received. And, probability of transition (b) is changed from 0 to 1 when symbol "end finger pointing" is received. Probabilities of transition (c) and (d) are changed in the same way. Both of transitions (c) are synchronized with one another. The same thing can be said of transition (d), (e) and (f).

Behavior model and motion model

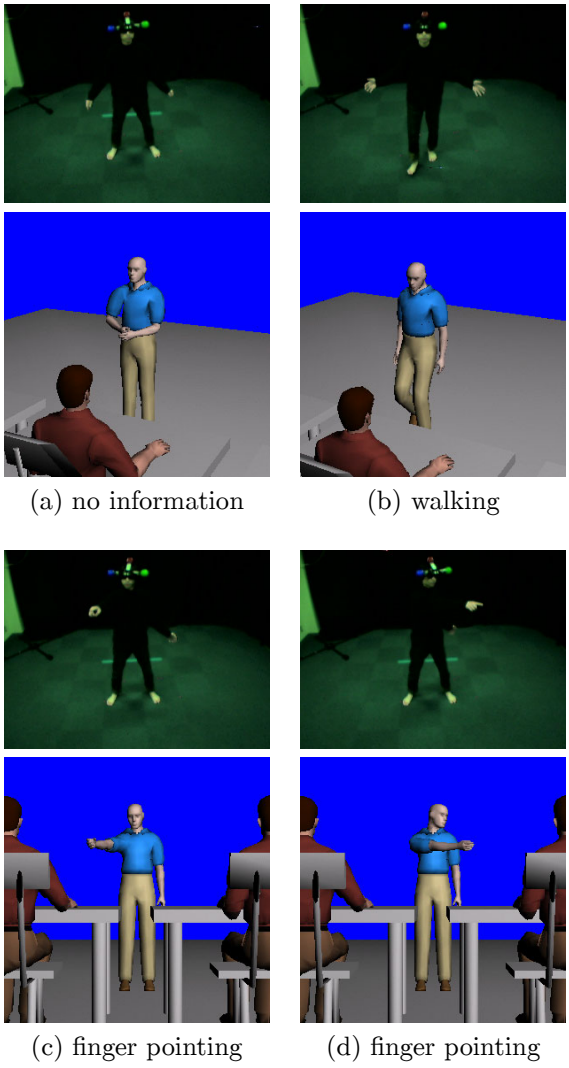
Reinforcing behavior model and motion model makes avatar's behavior more natural. Varying transition probabilities according to each participant individualizes avatar's behavior since the probabilities mean avatar's tendency of actions. In addition, we will research for automatic learning of the probabilities.

5. REFERENCES

[1] K. Russell, T. Starner, and A. Pentland, "Unencumbered virtual environments," in *Proc. of IJCAI'95 Workshop on Entertainment and AI/Alife*, Aug. 1995, pp. 58–62.

[2] M. Roussou, A. Johnson, T. Moher, J. Leigh, C. Vasilakis, and C. Barnes, "Learning and building together in a virtual world," *Presence*, vol. 8, no. 3, pp. 247–263, Jun. 1999.

[3] Satoshi Yonemoto and Rin-ichiro Taniguchi, "High-level human figure action control for vision-based real-time interaction," in *Fifth Asian Conference on Computer Vision*, Jan. 2002, pp. 400–405.



(a) no information (b) walking

(c) finger pointing (d) finger pointing

Fig. 3. Original and avatar images: (a) There is no information transmitted, and teacher's avatar is joining his/her hands which the system selects using behavior model. (b) A teacher and his/her avatar is walking (d) A teacher and his/her avatar are pointing to a student. (e) A teacher and his/her avatar are pointing to another student.

[4] Naoto Date, Hiromasa Yosimoto, Daisaku Arita, Satoshi Yonemoto, and Rin-ichiro Taniguchi, "Performance evaluation of vision-based real-time motion capture," in *CD-ROM Proc. of International Parallel and Distributed Processing Symposium*, Apr. 2003.