Journal of Universal Computer Science, vol. 13, no. 2 (2007), 161-176
submitted: 31/7/06, accepted: 15/1/07, appeared: 28/2/07 © *J.UCS*

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

Daisaku Arita
(Institute of Systems & Information Technologies/KYUSHU, Japan
arita@isit.or.jp)

Rin-ichiro Taniguchi
(Kyushu University, Japan
rin@limu.is.kyushu-u.ac.jp)

Abstract: We propose a concept of real-time human proxy for avatar-based communication 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 a simple game and a virtual classroom system. The experimental results shows us that RHP is useful for avatar-based communication.
Key Words: avatar-based communication, virtual reality, motion capture, motion generation
Category: H.5.1, H.5.2, H.5.3, I.3.7

1 Introduction

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

High quality of image Participants are taken by cameras, which produces image-based, or pixel-based data representation. It reflects all the appearance features of participants both of geometrical and photometrical ones, and can reproduce imagery with high quality.

Easy to use Audio-visual communication systems can be built without any image and sound processing. This means that we can use them easily only by connecting equipments such as cameras and displays.

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 communication 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 communication systems, there are several researches on virtual environments for human communication [Russell et al. 1995] [Roussou et al. 1999]. 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 communication 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 communication, which makes it easy to acquire and present participants' information.

2 Real-time Human Proxy

For natural avatar-based communication, 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 were present at distant places. RHP acquires verbal and nonverbal information, transfers acquired information of human activity, and presents transferred information in real-time.

In this paper, we focus on acquisition and presentation of nonverbal information. In the acquisition process, we symbolize the human action information under a given communication environment, such as classroom. In the presentation process, the symbols acquired are presented, or visualized, which are augmented based on the knowledge of the environment. [Fig. 1] shows a concept of RHP.

The important considerations behind the symbolization are summarized as follows:

– The important aspect of avatar-based communication is that an avatar, or an appearance of a human, can be changed depending on the purpose of communication,

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

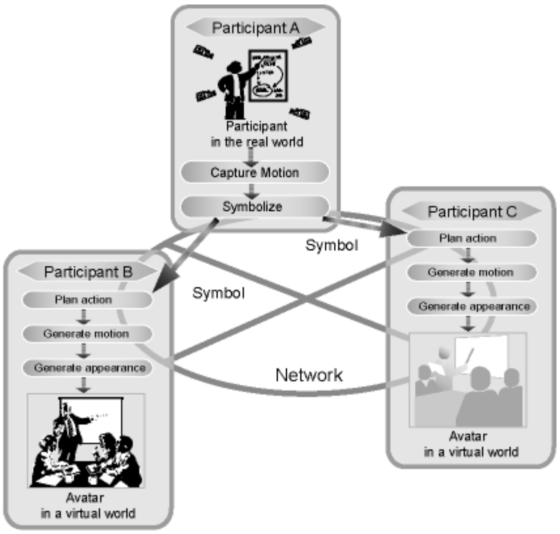


Figure 1: The concept of RHP.

attendance, etc. However, only with raw data of human motion, such as motion vectors of body parts, which are acquired by a motion capture system, only an avatar with the same physique as an observed human can be presented. It is quite difficult to present avatars with different physique or avatars with different body structure.

– By a motion capture system, very detailed motion information can not be extracted such as hand postures, face expression at the same time. Such details often express intention and are important for communication. Therefore, here, we interpret, with the aid of knowledge of the purpose communication, limited motion information into intentions of communication, i.e., *symbols*. In presenting the symbols, we can visualize an avatar so as to express the intentions efficiently, i.e., generate detailed motions which are not acquired by a motion capture system.

– The symbolization is also quite helpful to compress the amount of data transfer and to improve QoS (Quality of Service).

3 Information Acquisition

3.1 Motion Capture

To acquire nonverbal information, we use a real-time motion capture system which we are constructing[Date et al. 2004]. 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.

3.2 Symbolization

However, our motion capture system 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, transferring 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. And in case that a participant points to somewhere by his/her finger, recognizing the motion as finger pointing, transferring minimum information, and synthesizing motion parameters of arms and fingers generates more natural finger pointing scene.

For these reasons, the system categorizes motion sequences into pre-defined actions, expressing them as symbols. Each symbol is formed by a label of an action and its parameters. One example is symbol "walking (p_x, p_y, ν_x, ν_y)" where p_x and p_y are the position, ν_x and ν_y are the velocity of a participant. Another is symbol "finger pointing (P_n)" or symbol "finger pointing (d_x, d_y, d_z)" where P_n means a participant number pointed to and (d_x, d_y, d_z) means the direction of finger pointing.

After recognizing human actions from captured motion information, the system transfers the symbols to the representation side of a virtual space.

However, the rule which symbols and parameters to be transferred should be changed according to the situation. If all participants are seated, symbol "walking" is not necessary. If walking parameters are important for the communication such as rehabilitation and dance training, walking parameters such as angles of knees and ankles should be transferred in a raw data form.

The criterion for deciding which symbols and parameters should be transferred is intentionality of participants avatars, which depends on the situation. Then, we establish rules in advance describing which symbols and parameters should be transferred according to the situation of the communication.

4 Information Presentation

We define that an avatar is an object which is participant's substitute in a virtual space. An avatar has pre-defined knowledge to generate its motion and appearance from symbols. But it is time-consuming job to construct or modify the knowledge. Therefore the pre-defined knowledge is to be described in a reusable and extensible form. In addition, as described in [Section 2], RHP allows avatars to be designed beyond constraints of physical structure. To achieve these goals, we have employed a layered structure of the

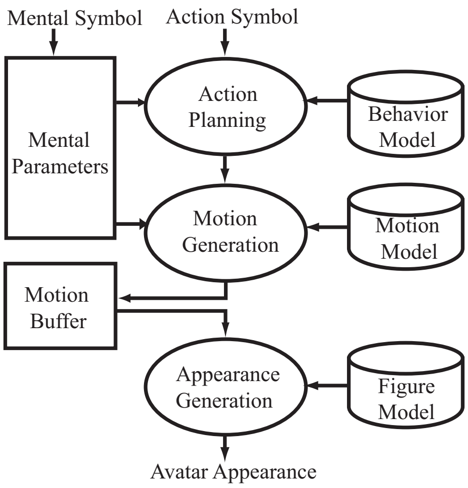


Figure 2: Process flows.

pre-defined knowledge. We divide the pre-defined knowledge into three layers (see Figure 2), and we make them independent as much as possible in order that we can easily modify physical structure of an avatar.

The three layers are *behavior model*, *motion model* and *figure model*. An avatar plans the next action based on *behavior model*, generates a motion corresponding to the next action based on *motion model*, and generates the avatar's appearance with motion based on *figure model*. Here, motion means posture sequences of an avatar's body parts.

4.1 Behavior Model and Action Planning

Action planner generates avatar's next action (action plan) such as "walking" and "raising hand" based on received *symbols*, *behavior model* and *mental parameters*[2]. Action plans are highly independent of avatar's physical structure. This allows model constructors to modify or replace *behavior model* with taking little care of relations between *behavior model* and other models.

4.1.1 Action Planning

A human can perform multiple actions at the same time if these actions do not require the same body part. For example, "walking (an action using right and left leg)" and

[2] The mental parameters are introduced for controlling action planning and motion generation. In addition they can be used for synchronizing nonverbal presentation and others such as verbal one and facial expression with each other. Since our current system uses the mental parameters rudimentarily, we do not discuss the mental parameters in this paper.

“raising hand (an action using right or left arm)” are not mutually exclusive. Therefore, the action planner should generate such multiple actions at the same time. Moreover, on RHP, symbols necessary for interaction depend on the kind of interaction. and it is desirable that *behavior model* can be modified easily, i.e., it should be as simple as possible. From the above considerations, the action planner plans an action referring to *behavior model*, which consists of two kinds of actions; (1)*outward action* is an action transiting from the neutral posture to a specified posture, (2)*Homeward action* is an action transiting from a specified posture to the neutral posture.

The neutral posture is the base posture of starting action. For instance of a human avatar, the posture is a standing posture with his/her arm taking down [see Fig. 3].

In general, the outward action can be planned in case that the posture of avatar’s body parts when a symbol is received is the same as the neutral posture. On the other hand, in case that the posture of avatar’s body parts when an action is planned is different from, or collides with, the neutral posture, the action can not be planned. However, if the collided posture is in the homeward action, then it can be planned, it is because avatar’s posture is to be the neutral posture soon. A homeward action can be planned after the corresponding outward action was planned.

An action is mainly planned according to a received symbol. However, an avatar often freezes if the avatar acts only when symbols are transmitted, since no symbols are transmitted when a participant does not make any pre-defined actions. Needless to say, such avatar’s behavior does not seem natural. To solve this problem, the action planner plans some actions spontaneously such as “folding arms” or “sticking hand into a pocket”, which have no influence on interaction. These actions are planned according to the mental parameters. Therefore, during no symbols are transmitted, an avatar can represent actions according to the participant’s mental state. To realize it, the system must understand the participant’s mental state correctly, which is one of our important future works.

4.1.2 Importance of Action

Each action has a degree of importance for realizing such a function that important actions, or actions according to symbols can be planned more preferentially than others. An example is given below in case when an outward action with a higher degree of importance is selected when an outward action with a lower degree of importance is presented. At first, the homeward action with a lower degree of importance is planned. Then, the outward action with a higher degree of importance is planned immediately. In the opposite case, an action with a lower degree of importance is ignored. Fundamentally, an action according to a symbol is given the highest importance, because the symbol explicitly presents an intention of the participant. On the other hand, an action unrelated to an interaction is given lower importance.

Figure 3: Neutral posture.

4.2 Motion Model and Motion Generation

There is a motion generator in an avatar which generates the motion based on the planned action, *motion model* and mental parameters. *Motion model* stores detailed motion information corresponding to each planned action.

4.2.1 Motion Generation

Motion model is represented as a table of correspondence between an action generated by the action planner and motion information which consists of the following information.

1. Keyframe sequence:$Q_1, Q_2, \cdots, Q_N$
2. The number of frames in the motion:M
3. Frame numbers of keyframes:$p_1, p_2, \cdots, p_N$
4. Interpolation function : $f(i)|i = 0, 1, \cdots, M$

The motion generator generates a motion, or posture sequence, corresponding to a received action. Keyframes expressed with Quaternions are key postures in a motion. Quaternion Q is defined using a rotation axis (V_x, V_y, V_z) and a angle θ as equation (1),

$$Q = (V_x \sin\frac{\theta}{2}, V_y \sin\frac{\theta}{2}, V_z \sin\frac{\theta}{2}, \cos\frac{\theta}{2}). \quad (1)$$

Then, a motion is generated by interpolating between keyframes by using of interpolation function $f(i)$ where i is a frame number in a motion. $f(i)$ is represented in a

Bezier function. The process of interpolation between Q_1 and Q_2 is described below. The difference between Q_1 and Q_2, called Q_{diff}, is calculated with equations (2), (3), (4) and (5),

$$Q = (x, y, z, w) \tag{2}$$
$$\bar{Q} = (-x, -y, -z, w) \tag{3}$$
$$Q_A Q_B = (v_A \times v_B + w_A v_B + w_B v_A, -v_A \cdot v_B + w_A w_B) \tag{4}$$
$$Q_{\text{diff}} = Q_2 \bar{Q}_1, \tag{5}$$

where $Q_A = (x_A, y_A, z_A, w_A) = (v_A, w_A)$ and $Q_B = (x_B, y_B, z_B, w_B) = (v_B, w_B)$. $(V_{x\,\text{diff}}, V_{y\,\text{diff}}, V_{z\,\text{diff}})$ and θ_{diff}, which are the rotation axis and the rotation angle between Q_1 and Q_2, can be calculated using equation(1). Using $(V_{x\,\text{diff}}, V_{y\,\text{diff}}, V_{z\,\text{diff}})$, θ_{diff} and $f(i)$, the motion $Q_{in}(i)$ $(p_1 \leq i \leq p_2)$ for moving from Q_1 to Q_2 is calculated with the equation(6),

$$Q_{in}(i) = (V_{x\,\text{diff}} \sin\frac{\theta_{in}}{2}, V_{y\,\text{diff}} \sin\frac{\theta_{in}}{2}, V_{z\,\text{diff}} \sin\frac{\theta_{in}}{2}, \cos\frac{\theta_{in}}{2}), \tag{6}$$
$$\theta_{in} = \theta_{\text{diff}} f(i). \tag{7}$$

In this example, we describe about a motion using only one body part. In the case of using multiple body parts, it is just to do these operation for every body part. And the motion generator change the interpolation function by moving control points according to the mental parameters. Therefore a motion can be changed according to the mental states at that time.

4.2.2 Motion Buffer

The motion buffer is a kind of queue. Basically, newly generated motions by the motion generator are added to the tail of the motion buffer, and the oldest motion at the head of the motion buffer is removed to generate avatar's appearance by the appearance generator. However, in case that a generated motion does not collide with an already stored motion, the generated motion is not added to the tail but unified motion is stored where the old stored motion was. This is because that these two motions can be represented simultaneously. For example, when a motion "walking (using right and left leg)" is already stored and a motion "pointing with right finger (using right arm)" is generated, these motions do not collide with each other. So they can be represented simultaneously, then they are unified and turn into one motion "walking pointing with right finger(using right and left leg and right arm)," which is stored where the motion "walking" was.

4.2.3 Motion fusion

As described before, we restrict actions planned by the action planner to only two kinds of actions which are outward actions and homeward actions. This can reduce animator's job, and make action planning simple. On the other hand, our aim is avatar-based interaction, so an action according to a symbol has to be represented immediately. Therefore such a system does not meet our aim that can not represent an outward action until a colliding homeward action has finished. Moreover, it is unnatural that all actions start from the neutral posture. To solve this problem, the motion generator fuses a homeward action and an outward action which collide with each other, in concrete, interpolates two motions.

The process of fusing two motions is described below. To make the explanation simple, both motions are single body part motions and the numbers of frames of both motions equal to M Motion of the homeward action is $Q_{h(i)}$ and motion of the outward action is $Q_{o(i)}$ $(0 \leq i \leq M-1)$. Difference between these motions, called $Q_{\mathrm{diff}(i)}$, is calculated with equation (5) using $Q_{h(i)}$ and $\bar{Q}_{o(i)}$. Using $Q_{\mathrm{diff}(i)}$, rotation axis $(V_{(i)_{x\,\mathrm{diff}}}, V_{(i)_{y\,\mathrm{diff}}}, V_{(i)_{z\,\mathrm{diff}}})$ and angle $\theta_{\mathrm{diff}(i)}$ for moving from $Q_{o(i)}$ to $Q_{h(i)}$ are calculated with equation(1), and the fused motion, called $Q_{c(i)}$, is calculated with equation(6) using an interpolation function $f(i)$. The interpolation function is important in order to fuse two motions smoothly. We have succeeded in obtaining results like Figure 4 using a linear function $f(i) = i/(M-1)$. In case of using multiple body parts, it is just to do these operations for every body part.

4.3 Figure Model and Appearance generation

Figure model stores avatar's geometry data and physical structure, various of which are provided as computer graphics characters in a lot of computer graphics commercial softwares. We can select as an avatar not only a realistic human body but also a human-like-structured body such as a cartoon human, a humanoid robot or a lion and moreover anything we want to use such as a bird, a fish or a car. In case of using a human-unlike-structured body, we have to modify the body for easy communication. For example, Nemo, an anemone fish, has eyebrows and white of the eye. This is not realistic but useful for expressing his attention and emotion.

The appearance generator generates avatar's appearance using the posture from the head of the motion buffer and *Figure model*.

5 Prototype System of RHP

We have developed two interaction systems to verify the effectiveness of RHP.

5.1 Simple game

In this experiment which is a simple interactive game, we verify the effectiveness of RHP in case that all actions to be acquired and presented for interaction are able to be enumerated.

The rules of the game, which is a simplified version of a famous game in Japan, are as follows:

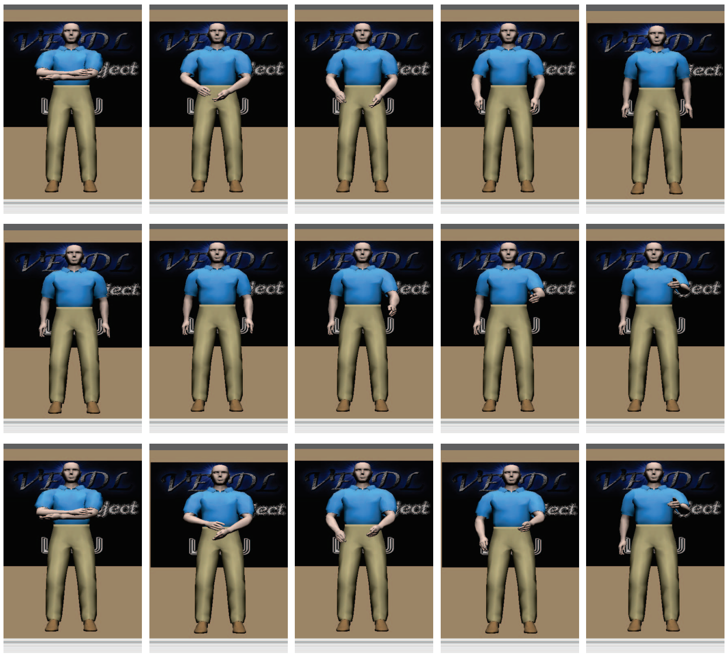
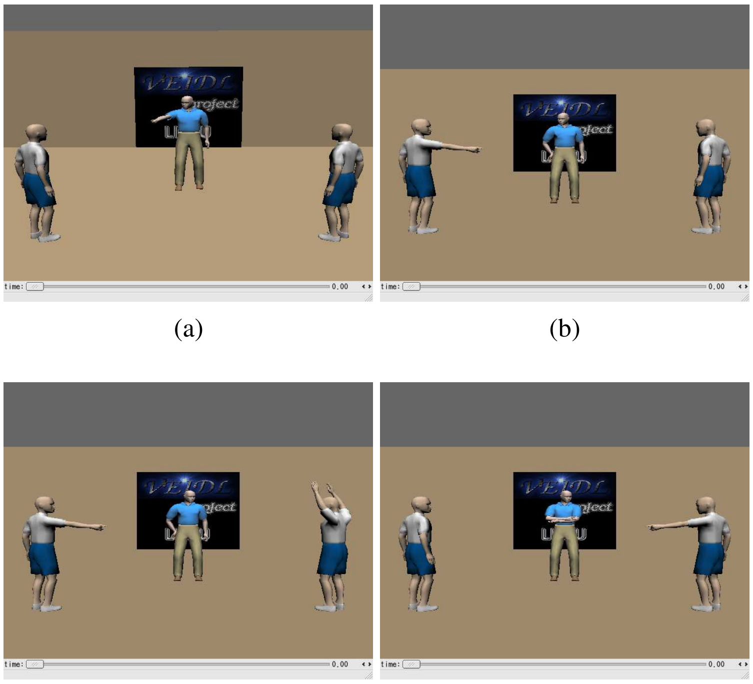


Figure 4: Fusing motion: The motion of a homeward action, that of an outward action and the fused motion are shown in the top row, the middle one and the bottom one respectively.

1. One of participants becomes a leader.
2. The leader says “A” and points to another participant.
3. The participant who is pointed to at step 2 says “B” and points to another participant.
4. The participant who is pointed to at step 3 says “C” and puts his/her hands up.
5. The participant who puts his/her hands up becomes a leader.
6. Return to step 2 until someone fails.

The labels of actions which are parts of pre-defined knowledge are as follows:

– Outward of *finger pointing*, and homeward of it: actions activated by a symbol.
– Outward of *hands up*, and homeward of it: actions of activated by a symbol.
– Outward of *head turn*, and homeward of it: actions of activated by a symbol.



(a) (b)

(c) (d)

Figure 5: Snapshots of the game. (a) A leader points to a participant. (b) A participant points to another participant. (c) A participant puts his/her hands up. (d) A participant becomes a leader and the next turn starts.

- Outward of *folding arms*, and homeward of it.
- Outward of *putting hand to waist*, and homeward of it.

Some snapshots of the game are shown in Figure5.
The impressions of participants about representation as follows:

- An avatar can represent pre-defined actions which participants behaved. And it can present actions which does not indicate participant's intentions during the participant does not make any pre-defined actions. Therefore avatar's behavior is natural.
- When a participant acts *finger pointing* during its avatar acts *arms folding*, the avatar immediately finishes folding arms and starts finger pointing.
- Participants are able to easily understand where avatars looks and points.

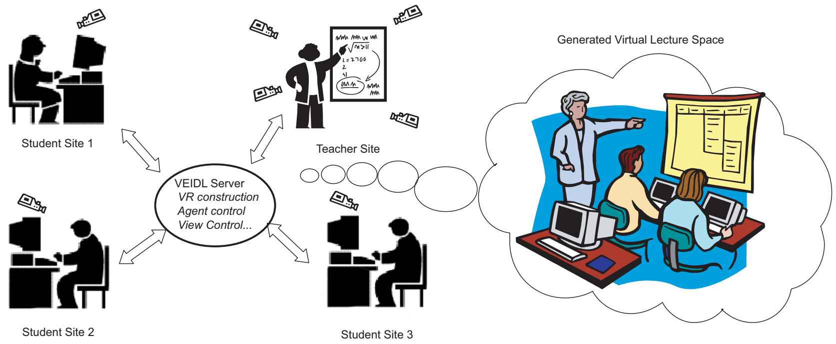


Figure 6: The concept of VEIDL.

5.2 VEIDL

We are also 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. 6.

In this experiment, there are one teacher and three students in a virtual classroom leaning English conversation with role playing. Transferred nonverbal information, or symbols, are finger pointing, hand raising and head moving. The story of role playing is displayed on a screen of the virtual classroom [Fig 7]. First, the teacher assigns roles of A, B and C to three students. Then, students start playing their roles.

The overview of the virtual classroom is shown in [Fig. 8 and Fig. 9]. Transferred data size is reduced to about 0.35% in comparison with the motion capture data. The results of this experiment shows us the advantage same as the first experiment. However, there are some disadvantages as follows:

- Some delay occurs between participant's action and avatar's action caused by motion recognition.
- It is necessary to synchronize verbal and nonverbal information presentation[3]

6 Conclusions

In this paper, we propose a concept of real-time human proxy for avatar-based communication systems, which virtualizes a human in the real world in real-time and which lets

[3] Since all participants are in a room with partition board, verbal information is directly reached to each other. Then, there is no synchronization between verbal and nonverbal information presentation.

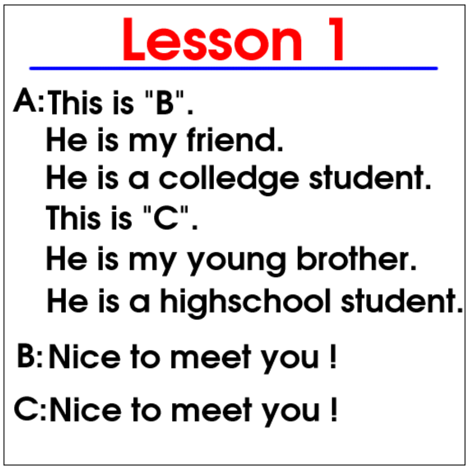


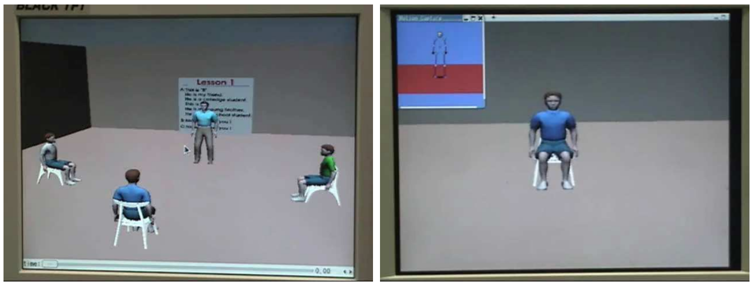
Figure 7: Story of role playing.

the virtualized human behave as if he/she was present at a distant place. For estimating RHP, we apply it to a simple interactive game and VEIDL which is a virtual classroom system. The experimental results shows us that RHP is useful for avatar-based communication.

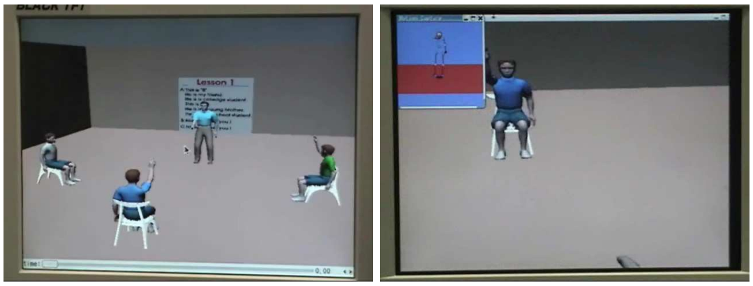
There are a lot of future works as follows:

Which actions should be transferred as symbols? For RHP, system constructors have to enumerate actions to be transferred according to the communication situation. However, it is very difficult to make a complete list necessary for the communication situation. For relieving this difficulty, we are now researching for semiautomatic enumeration method, which allow system constructors to manually select actions to be transferred among actions automatically extracted from motion capture data of a participant in the communication situation [Araki et al. 2006].

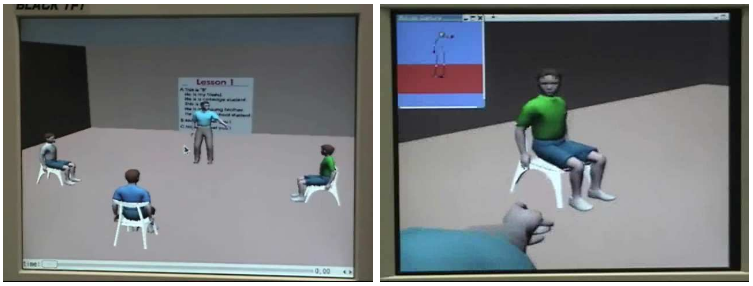
It is time-consuming to construct motion model Behavior model can be shared with various physique of avatars. On the other hand, motion model can not be shared. This means that introducing a new avatar requires a hard and time-consuming task of preparing motion parameter sequences corresponding to all actions not only to be transferred but also spontaneously planned by the action planner. To solve this problem we are researching for automatic generation of motion parameter sequences by using the motion retargeting technique [Gleicher 1998], which can adjust motion parameter sequences of a base avatar to a new one which has different body size. In addition, we are now researching for real-time motion generation using motion model, which can allow body parts to avoid conflicting with other body parts and surrounding objects.



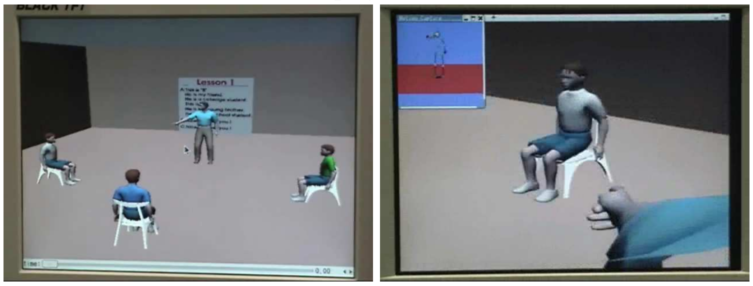
Teacher: Does anybody want to role A



Teacher: O.K. Please play the role of A, Mr. Blue.

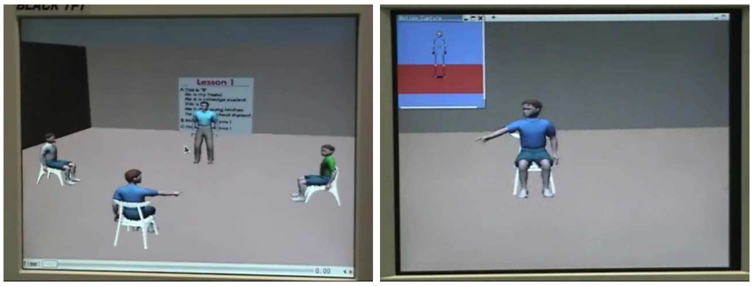


Teacher: O.K. Please play the role of B, Mr. Green.

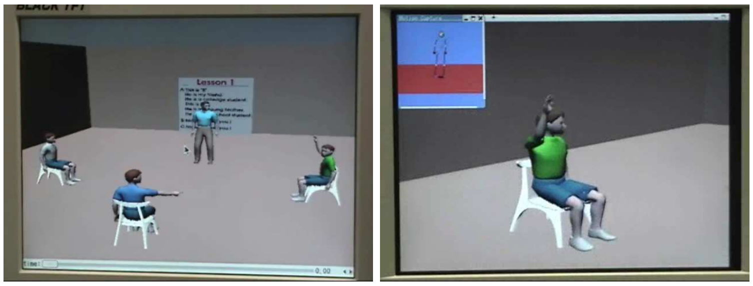


Teacher: O.K. Please play the role of C, Mr. White.

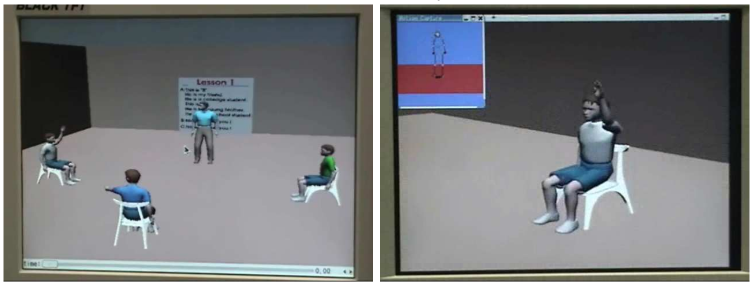
Figure 8: Virtual classroom. Left: Overview of classroom. Right: Teacher's view.



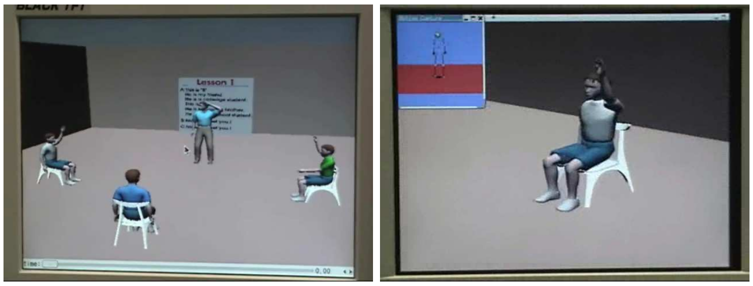
Student(Blue): “This is Mr. Green.”



Student(Blue): “He is my friend. · · ·”



Student(Blue): “This is Mr. White. · · ·”



Students(Green, White): “Nice to meet you !”

Figure 9: Virtual classroom (Cont.).

How to reduce influence of delay In RHP communication, there is some delay between participant's action and avatar's action caused by action recognition and network transfer. We are now researching for reducing delay caused by action recognition by early recognition technique [Mori et.al. 2006].

How to acquire and express mental status However we currently have no clear answer how to acquire and express mental status, mental status is very important for RHP communication. For acquisition we need to introduce not only a motion capture system but also facial expression recognition system. For expression we can introduce a lot of knowhow to make animation movies such as "Finding Nemo", "Cars" and so on.

Acknowledgments

This work has been partly supported by "Intelligent Media Technology for Supporting Natural Communication between People" project (13GS0003, Grant-in-Aid for Creative Scientific Research, the Japan Society for the Promotion of Science) and "Real-time Human Proxy for Avatar-based Distant Communication" (16700108, Grant-in-Aid for Young Scientists, the Japan Society for the Promotion of Science).

References

[Russell et al. 1995] Russell, K., Starner, T., Pentland, A.: "Unencumbered Virtual Environments"; Proc. of IJCAI'95 Workshop on Entertainment and AI/Alife (Aug 1995), 58–62.

[Roussou et al. 1999] Roussou, M., Johnson, A. Moher, T., Leigh, J., Vasilakis, C., Barnes, C.: "Learning and Building Together in a Virtual World"; Presence, 8, 3 (Jun 1999), 247–263.

[Date et al. 2004] Date, N., Yoshimoto, H., Arita, D., Taniguchi, R.: "Real-time Human Motion Sensing based on Vision-based Inverse Kinematics for Interactive Applications"; Proc. of International Conference on Pattern Recognition, 3 (Aug 2004), 318–321.

[Araki et al. 2006] Araki, Y., Arita, D., Taniguchi, R., Uchida, U., Kurazume, R., Hasegawa, T.: "Construction of Symbolic Representation from Human Motion Information"; Proc. of International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (Oct 2006).

[Gleicher 1998] Gleicher, M.: "Retargetting Motion to New Characters"; Proc. of SIGGRAPH (Jul 1998), 33–42.

[Mori et.al. 2006] Mori, A., Uchida, S., Kurazume, R., Taniguchi, R., Hasegawa, T., Sakoe, H.: "Early Recognition and Prediction of Gestures"; Proc. of International Conference on Pattern Recognition (Aug 2006), Tue-P-II-3.