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

Kunita, Masashi
Department of Intelligent Systems, Kyushu University

Yoshimatsu, Hisato
Department of Intelligent Systems, Kyushu University

Hayama, Daisuke
Department of Intelligent Systems, Kyushu University

Arita, Daisaku
Department of Intelligent Systems, Kyushu University

他

https://hdl.handle.net/2324/5923

出版情報：Proceedings of the 10th Korea-Japan Joint Workshop on Frontiers of Computer Vision, pp.71-76, 2004-02
バージョン：
権利関係：

REAL-TIME HUMAN PROXY: AN AVATER BASED INTERACTION SYSTEM

Masashi KUNITA, Hisato YOSHIMATSU, Daisuke HAYAMA,
Daisaku ARITA, Rin-ichiro TANIGUCHI

Department of Intelligent Systems, Kyushu University, Japan
{mkunita, hisato, daizz, arita, rin}@limu.is.kyushu-u.ac.jp

Abstract This paper describes techniques for improving human representation on an avatar-based interaction system, using a motion capture system and human action symbolization. Avatar-based interaction systems with computer-generated virtual environments have difficulties in acquiring user's information, which is enough to represent the user as if he/she were in the environment. This mainly comes of high degrees of freedom of human body. Since it is almost impossible to acquire all information of human motion, we turn acquired motion sequences into symbols, which show what kind of action has occurred, by recognizing human action with rough motion capturing. In this paper, we describe our approach on both side of acquiring and representing human action, and finally, experimental results are presented.

1. Introduction

Communication technologies have made it usual to have communication for people who are distributed. For example, telephone is one of the most useful tools to communicate using acoustic information, and e-mail is one of them using literal information. Furthermore, richer tools are also available. Video chat or videophone has been made possible by higher data compression techniques and broader network bandwidth. Many people may consider videophone to be the richest communication tool, since videophone seems to provide us the feeling of being connected to other speaker's place, because appearances of speakers are directly transported to each other. It has problems, however, in case that many (10 or more) people are participating at once, as on a conference or a meeting. Every participant in a conference must have numbers of windows showing the other participants on his/her display. Each window collocated on the same plane has its own geometrical coordinate system, showing his/her companion who is always just facing the front. This does not make participants feel as if they were present alongside. To solve this problem, we chose the way to get all participants into a virtual space.

There are several researches on virtual environments for human interaction. In these researches, a 3-D virtual space is reconstructed, in which each participant is represented as an avatar by computer graphics techniques. Through the reconstructed virtual space, each participant sees and hears other participants' activities from the position where his/her avatar is represented. Their positional relations are consistent virtually. This means that each participant can understand where other participants stand, look at, and point to, or understand where a sound comes from, and also move around in the virtual space. In contrast to a video chat system, participants can easily feel coexistence. However, these avatar-based interaction systems have difficulty in controlling avatars, where the degrees of freedom of human body are so high that legacy input devices are not sufficient to acquire or input participants' activities.

In this paper, we describe Real-time Human Proxy; our concept for improving human representation in avatar-based interaction.

2. Real-time Human Proxy

2.1. Use of Motion Capture System

In avatar-based interaction, an avatar is expected to reflect activities of a participant into a virtual space as if he/she were there. An avatar is expected to turn its head to a particular person when a participant does so, for example. Nevertheless, as already mentioned above, legacy input devices are not sufficient to acquire participants' activities in aspects of quality and quantity. Using such devices, participants have to keep feeding their own activities into a system by hand, and acquired information may not be precise. Special input devices are often used as solution to this problem[1]. We use a motion capture system (MCS) as an input device, for acquiring more information of participants without compelling them annoying operations.

2.2. What Real-time Human Proxy is

Although using an MCS as an input device, it is not possible to reflect all of participant's motions into an avatar. More information could be acquired in comparison with legacy devices, but an MCS cannot acquire all information such as angles of fingers. Just pouring captured data into an avatar, the lack of information may cause unnatural avatar motions. On the other hand, it is not necessary for an avatar to act just as same as participant's motions, since participants usually want to know only what others are doing, but how others moving. Real-time Human Proxy (RHP) is a new concept for avatar-based interaction, which makes better use of an MCS, and makes avatar act more meaningfully.

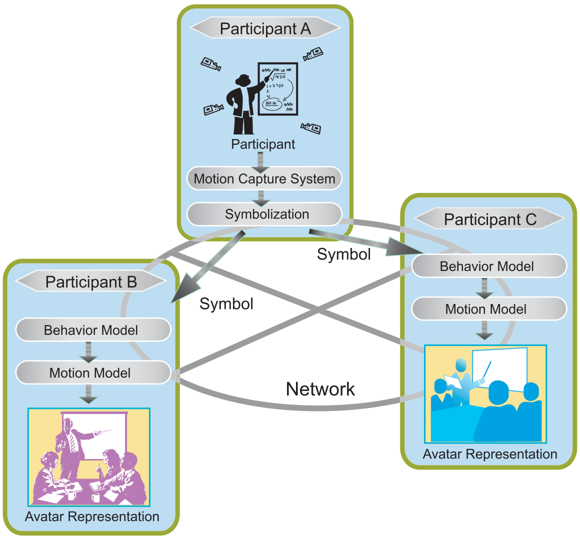

Figure 1: This figure shows how an action of a participant A take effect to sights of participant B and C on RHP. Transported symbols are changed into motion sequences on each site. In reality, any other participant also sends symbols simultaneously.

RHP is a concept which virtualizes a human in the real world in real-time. The aim is to make an avatar act as if he/she in a distant place is present in a virtual space. Therefore, we focus on acquisition and representation of human action or nonverbal information. We use *symbolization* on acquisition, and use *motion model* and *behavior model* on representation, that they are illustrated in figure 1 and described below.

2.3. Symbolization

On RHP, we acquire human actions instead of human motions. We categorize motion sequences into pre-defined actions, expressing them as *symbols*. Each symbol is formed by a label of an action and its parameters, such as 'walking (v_x, v_y)' where (v_x, v_y) is velocity of participant. After recognizing human actions from captured motion data, the system transports the symbols to the representation side of the virtual space.

2.4. Motion Model

Since each symbol transported from a participant is based on a label of an action, to make an avatar act, therefore, pre-defined knowledge is needed, which is a set of trajectories or motion sequences of each body part, tied to each symbol. We call the knowledge *motion model*.

2.5. Behavior Model

We have another model called *behavior model*. Behavior model is a set of state transition graphs that decides the next action an avatar is going to do. Each state in the graphs corresponds to an action or a symbol such as 'walking', 'raising hand', and 'pointing with finger.' When a symbol is transported, the current state of the graph is forced to transit to the state corresponding to the symbol.

In addition, an avatar often freezes if the avatar acts only when symbols are transported, since no symbols are transported 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, behavior model has some actions invoked spontaneously such as 'folding arms' or 'sticking hand into a pocket.' These actions may work for time filling, and may make participants feel more natural[2]. And of course, these actions do not indicate participant's intentions in order not to influence interaction between participants.

2.6. Representation of Virtual Space

Using these models, avatars behave in a virtual space according to transported symbols. At first, a behavior model changes transition probabilities in its transition graphs, encouraging states according to the symbol that is just received. Secondly, an avatar performs an action referring motion models tied to the state just have been activated in the behavior model.

A participant is able to see a virtual space in which any participant, including him/herself, is represented as an avatar. Therefore, the viewpoint is anchored where the avatar was represented.

2.7. Benefits

The use of an MCS liberates participants from annoying operations of controlling an avatar. On the symbolization process, the system recognizes the action that a participant makes. This operation is intuitive and natural since participants make the same action as they make in the real world.

The behavior model can prove consistency in transition of action. Unnatural transition of action (or a behavior) will not be provided, where any transition from an action to another is a production of transition probabilities in the behavior model, and it affects as a role of the rule. That means impossible transition of action is preventable.

The concept, i.e. symbolization of actions, and representation of an avatar using motion model and behavior model, lets representation process simple. As described above, acquired information about participant's motion has lacks of several body parts, such as angles of fingers, which must be compensated proving conformity with other body parts to provide natural representation. Using RHP, actions are already recognized, and it is possible to generate full body motion sequence in an arbitrary manner referring the motion model.

Moreover, this allows avatars to be designed beyond constraints of physical structure. Ordinary, each avatar is designed to fit an actor whose motions are captured, so that 3-dimensional positions of body parts acquired from an MCS is correctly represented in computer graphics. Some errors in interpretation of posing occur from mismatches in size or length of any body parts, such as height of body or lengths of arms. RHP doesn't transport the 3-dimensional positions directly, and consequently eliminates the constraints. We can use any kind of avatar including higher body, shorter arms, bigger head, etc. This improves not only usability of the system, but also variety of avatars in interaction.

3. Prototype System of RHP

We are developing a system called VEIDL (Virtual Environment for Immersive Distributed Learning), which is a prototype system for estimating RHP. VEIDL is a virtual classroom environment where geographically dispersed people can attend through avatars of teachers or students represented, as shown in figure 2.

3.1. System Overview

On VEIDL, every participant has his/her own microphone and cameras as input devices, which acquire verbal and nonverbal information from him/her, and a display unit (possibly HMD) and speakers as output devices, which present a scene of the classroom generated by a computer. Acquired information from each site is transported via network to one another. Each participant can see a computer-generated scene of the classroom from the viewpoint of his/her avatar, and, of course, can also see the other participants behaving according to the transported information. This makes participants able to interact each other

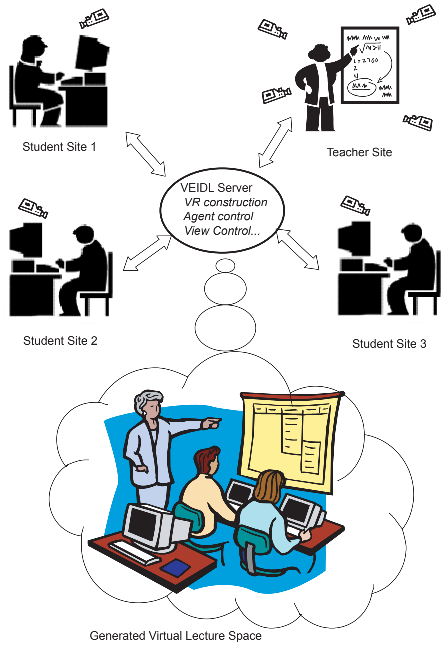

Figure 2: The Concept of VEIDL

through the virtual environment.

The advantage of dealing with a virtual classroom, as a prototype of RHP, is that it is easy to decide which information (or symbol) should be transported, since the objective of interaction in a classroom is clear.

The number of participants is currently limited up to seven. This is because VEIDL is just for the purpose of estimation of RHP, where every participant is expected to interact to each other and not to only spectate.

3.2. Capturing Action

To acquire nonverbal information, we use two types of real-time MCS, which we are constructing. One is for full-body motion capturing, and the other is for upper-body motion capturing. They are described more precisely below.

The full-body MCS[3] uses ten cameras around a human and a PC cluster with RPV, which is a programming environment for real-time parallel image processing. We can get 3-dimensional positions of a head, hands, elbows, knees, feet, and a torso in real-time without any markers attached with body parts.

The upper-body MCS[4] uses two cameras in front of a human. This system is designed for desktop interface. We can get 3-dimensiotnal positions of a head and both hands in real-time without any markers.

On VEIDL, two types of participants are conceivable, a teacher and students. We assume that the full-body MCS may be applied to a teacher, and the upper-body MCS may be applied to each student.

Additionally (if possible), we use a HMD with markers for determining direction of a face (shown in figure 3). This makes it possible to provide us an immersive view of a virtual space.

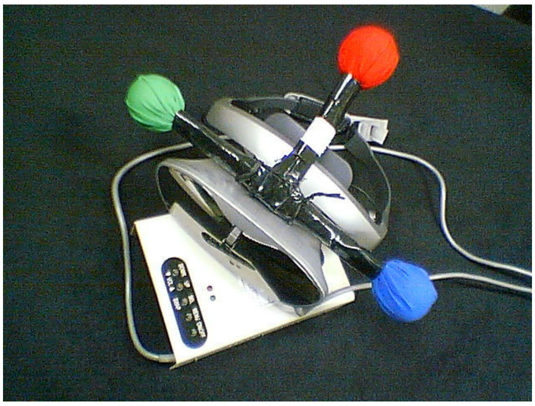

Figure 3: A Marker-Attached HMD

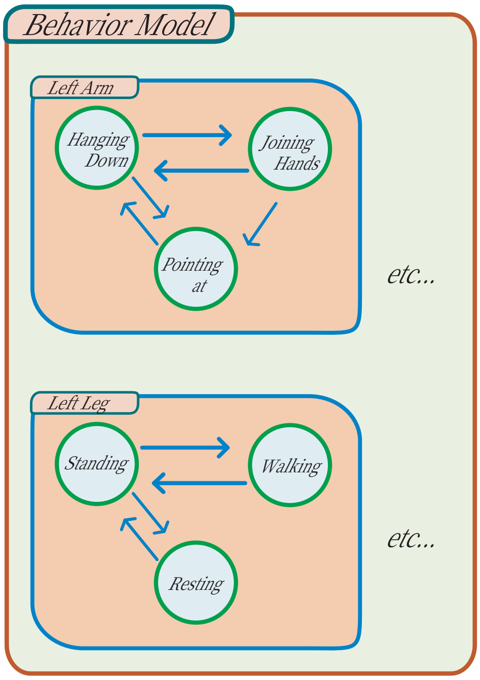

Figure 4: Behavior Model (partial)

We have decided to have four kinds of basic nonverbal information or symbols that seems strongly necessary for presenting teacher's intentions. They are 'walking', 'turning body', 'looking at', and 'pointing with finger at.' We use simple technique[5] for action recognition where we have only a few symbols to recognize at the present stage.

3.3. Transporting

Symbols from each participant are transported to other participants via network. For the characteristics of this application, we use Real-time Transport Protocol[6] as a base protocol, which is real-time oriented and supports multi-casting. Symbols from each participant are sent to and received from other participants concurrently. In addition, audio streams containing participants' voice are transported in the same manner.

3.4. Representing Action

We defined a behavior model that consists from five state transition graphs as partially shown in figure 4. Each graph inside the model corresponds to a body part of right and left arm, right and left leg, and a head. Transition probabilities between the states are defined manually. Motion models are also defined for each state in the behavior model. They are briefly defined manually, but will be replaced with more realistic model to improve reality of motion using human motion parameters observed from precise motion capture system and a hand shape measuring system.

We arranged two actions for the time filling, i.e. ‘joining his/her hands’ and ‘standing at ease’. They are invoked spontaneously by the behavior model according to transition probabilities in it, when no symbols are transported.

Using the MCS with marker-attached HMD, a participant can see around just panning and tilting their face as in the real world, for direction of a face is acquired clearly.

4. Experiment and Results

We had an experiment using the prototype system to estimate the concept, especially from the aspect of effect to avatar manipulation and perceptional relevance of appearances on avatar representation.

4.1. Environment

VEIDL was attended by five participants. They consisted of a teacher and four students. We focused on single participant for the objective of this environment, hence only the teacher avatar was controlled with the full-body MCS by a real human. The other avatars of students were controlled by an experimenter as dummy participants.

Subjects played role of a teacher in the MCS wearing marker-attached HMD. They are asked to make actions of registered and non-registered symbols. Appearances of subjects in both of the real world and the virtual space are observed by an experimenter with the symbols transported from the teacher’s side of the system.

4.2. Results

Avatar Manipulation

The results are shown in figure 5. Registered actions that subject was performed were correctly reflected in the avatar representation. Avatars could be manipulated intuitively.

Appearance of Motion

According to the received symbols, the avatar that is represented in the virtual space behaved smoothly. It is also true in the point of changing action such as from ‘joining his/her hands’ to ‘pointing with finger at’.

Discussion

RHP took a good affect since it is difficult to create precise motion compensating the lack of information, such as articular angles of wrists and twist angles of shoulders or hip joint, only using such sparse data of each body parts from the MCS we used.

The motions of the avatar were awkward since motion models had been created manually. This problem will be solved by reinforcing motion models by utilizing MCSs for getting reference motion parameters. Additionally, avatar behavior is limited yet. We need more symbols to transport for improving the quality of avatar expression.

5. Conclusion

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 were present. For estimating RHP, we apply it to VEIDL, which is a virtual classroom system. The experimental results show us that RHP is useful for avatar-based interaction.

References

[1] 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.

[2] Daisuke Hayama, Takehiko Hanada, Hiromasa Yoshimoto, Daisaku Arita, and Rin-ichiro Taniguchi, “Realization of a teacher’s avatar behaving naturally for Immersive distributed learning,” Proceedings of the 2003 IEICE general conference, p. 264, Mar. 2003.

[3] N.Date, H.Yoshimoto, D.Arita, S.Yonemoto and R.Taniguchi, “Performance Evaluation of Vision-based Real-time Motion Capture,” Proc. Workshop on Parallel and Distributed Computing in Image Processing, Video Processing, and Multimedia, in IPDPS CD-Rom Proceedings, 2003.

[4] Satoshi Yonemoto and Rin-ichiro Taniguchi: “A Direct Manipulation Interface with Vision-basedHuman Figure Control”, Proc. HCI International 2003, pp.811-815, 2003.

[5] Hisato Yoshimatsu, Daisaku Arita, and Rin-ichiro Taniguchi, “Symbolization of nonverbal information about a teacher for Immersive distance learning,” Proceedings of the 2003 IEICE general conference, p. 265, Mar. 2003.

[6] http://www.cs.columbia.edu/~hgs/rtp/

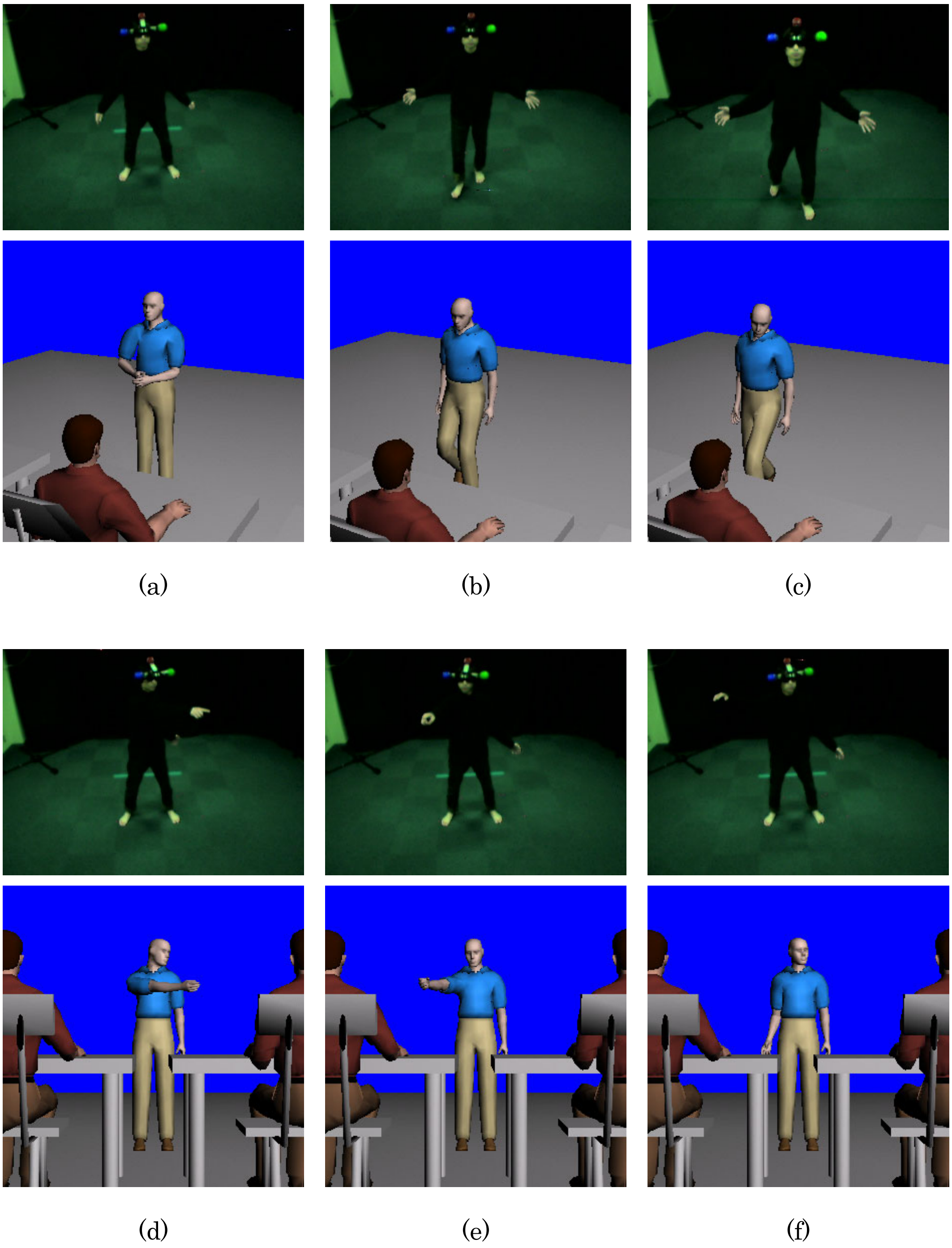

Figure 5: Appearance of a participant in the real world (above) and in a virtual space (below). (a) There is no symbol associated with the participant's action, and an action 'joining his hands' is invoked by the behavior model. (b)(c) The participant is walking. (d)(e) The participant is pointing at a student. (f) The participant is looking at a student.