LETTER

Embedding the awareness state and response state in an image-based avatar to start natural user interaction

Tsubasa MIYAUCHI[†], **Ayato ONO**[†], **Hiroki YOSHIMURA**[†], **Masashi NISHIYAMA**[†a)], ***and*** **Yoshio IWAI**[†b)],

SUMMARY We propose a method for embedding the awareness state and response state in an image-based avatar to smoothly and automatically start an interaction with a user. When both states are not embedded, the image-based avatar can become non-responsive or slow to respond. To consider the beginning of an interaction, we observed the behaviors between a user and receptionist in an information center. Our method replayed the behaviors of the receptionist at appropriate times in each state of the image-based avatar. Experimental results demonstrate that, at the beginning of the interaction, our method for embedding the awareness state and response state increased subjective scores more than not embedding the states.

key words: *Image-based avatar, Beginning of interaction, Awareness state, Response state*

1. Introduction

There is a demand for a system that enables a human and machine to naturally interact through an avatar [1]–[5]. Such a system would have the ability to interact with users in various scenarios, for example, navigating in public spaces (e.g., airport, station, or city office) throughout 24 hours. We discuss an image-based avatar system [3]–[5] in which a video sequence acquired from a real person is used on a life-sized display. We consider that the image-based avatar waits for a user and starts to navigate toward the user in an information center. In particular, we focus on the beginning of the interaction until the image-based avatar is asked a question regarding the location to which the user wishes to go.

The image-based avatar requires wait, awareness, response, and action states to smoothly start the interaction with the user. In the wait state, the image-based avatar is on standby because there are no users nearby. In the awareness state, the image-based avatar notices that the user is approaching, but because the distance from the user is still substantial, the avatar prepares to respond to the user at any time. In the response state, the image-based avatar returns a behavior to indicate that it is ready to accept the interaction with the user. In the action state, the image-based avatar talks with the user and performs a close interaction. In particular, the awareness state and response state play important roles to smoothly start the interaction. If the response state is not embedded in the image-based avatar, it is difficult for the user to decide when to start the interaction because the avatar is non-responsive despite the user approaching. Furthermore, if the awareness state is not embedded in the image-based avatar, the approaching user will be confused because the timing of the response of the avatar is delayed. The existing methods for image-based avatars mainly focus on the action state [3], [4] and partly on the wait state [5]. However, these existing methods do not sufficiently consider how to address the awareness state and response state for the image-based avatar.

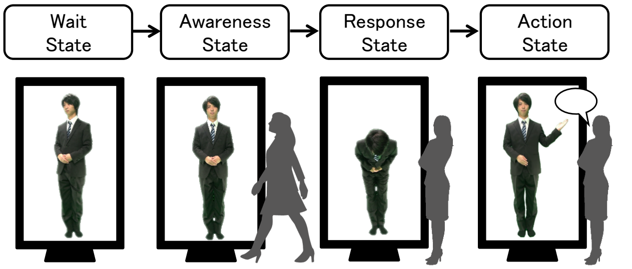

Fig. 1 The image-based avatar requires the above four states to smoothly start the interaction with the user.

In this paper, we propose a novel method for embedding the awareness state and response state in the image-based avatar to smoothly start an interaction with a user. We observe the beginning of an interaction between a user and receptionist in an information center. Based on the observation, we design a model of the receptionist's behaviors. Our method replays the behaviors of the receptionist when starting an interaction. We synthesize a video sequence of the image-based avatar to adjust the movements of the user. The results of a subjective assessment demonstrate that an avatar with the awareness state and response state obtains higher scores than an avatar without the states.

2. Design of the model for starting the interaction

2.1 Observation of receptionist

We observed an interaction between a user and receptionist in an information center. In many situations, the user asked the receptionist for the location of a place, such as a ticket counter or boarding gate. The receptionist performed passive behaviors until she was spoken to by the user. The receptionist was spoken to by the user infrequently. We interviewed the receptionist after the observation to design the model of the receptionist's behaviors. The details of the

Manuscript received January 1, 2015.
Manuscript revised January 1, 2015.

[†]Graduate School of Engineering Department of Information and Electronics, Tottori University, Tottori 680–8550, Japan
a) E-mail: nishiyama@eecs.tottori-u.ac.jp
b) E-mail: iwai@eecs.tottori-u.ac.jp

DOI: 10.1587/transinf.E0.D.1

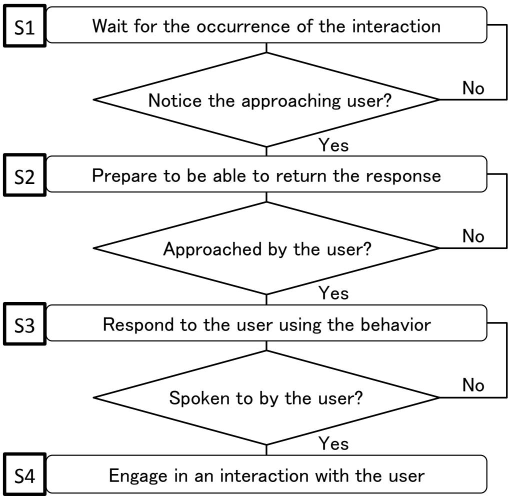
S1
Wait for the occurrence of the interaction
Notice the approaching user?
No
Yes
S2
Prepare to be able to return the response
Approached by the user?
No
Yes
S3
Respond to the user using the behavior
Spoken to by the user?
No
Yes
S4
Engage in an interaction with the user

Fig. 2 Model of the receptionist's behaviors when starting an interaction with a user.

model are described below.

2.2 Model of the receptionist's behaviors

Figure 2 shows the model of the receptionist's behaviors when starting an interaction with the user. In S1, the receptionist waits for the occurrence of the interaction because there was no user near to the receptionist. The receptionist's behavior involves looking forward or looking at the surroundings. In S2, the receptionist notices the approaching user and prepares to return a response to the user. The receptionist's behavior involves turning his/her face toward the user. In S3, the receptionist responds to the user to create an atmosphere in which it is easy to be spoken to. The receptionist's behavior involves looking at the user or bowing to the user. In S4, the receptionist is engaged in a close interaction with the user, such as a conversation about a location. We asked why the receptionist performed the aforementioned behaviors. The following situation was reported: The receptionist waited until the user approached before starting the interaction because a conversation was difficult from a distance in a public space. Thus, our model uses the distance between the user and receptionist to determine the transitions from S1 to S2, and S2 to S3. When the receptionist was approached by the user, the receptionist was asked for a location. The user first said 'Excuse me' and then asked questions. Our model uses the speech utterance by the user to the receptionist to determine the transition from S3 to S4.

2.3 Our method for replaying the receptionist's behaviors

In the following, we design our method for replaying the receptionist's behaviors using the image-based avatar. We describe each state of the receptionist shown in Figure 2 as follows: action state (S1), awareness state (S2), response state (S3), and action state (S4). The avatar performs the behaviors as follows: looking forward or looking at the surroundings in S1, looking at the user or turning its face toward the user in S2, looking at the user or bowing to the user in S3, and speaking and hearing in S4. The details of our method for replaying the behaviors are as follows: When the distance d_t from the user to the avatar is threshold D_1 or less, the state transition from S1 to S2 occurs. When d_t is D_2 or less, the state transition from S2 to S3 occurs. When the average a_t of the audio powers, which are acquired from the direction of the user's standing position, in the short term is threshold A or greater, the state transition from S3 to S4 occurs. In advance, we determine D_1, D_2, A by measuring the distance or audio power when transitioning from one state to another. Our method in real time measures d_t, a_t using a depth sensor and microphone.

3. Synthesizing the video sequence of the image-based avatar to replay the behaviors

3.1 Video clips to represent the behaviors

The image-based avatar is represented by a video sequence acquired from a real person. Thus, it is time-consuming to shoot all combinations of the receptionist's behaviors in advance. Instead of acquiring a video sequence, we consider acquiring video clips to represent the behaviors in S1, S2, and S3. Note that we do not address how to represent the behaviors in S4 because we focus on the beginning of the interaction. The video clip consists of the following: maintaining the initial posture in which the avatar faces the front, making a movement, and returning to the initial posture. Our method uses the following video clips:

- **C1**: Keep looking forward in the initial posture.
- **C2**: Turn left from the initial posture and return to the initial posture.
- **C3**: Turn right from the initial posture and return to the initial posture.
- **C4**: Bow from the initial posture and return to the initial posture.

A facial image of a frontal direction on a flat display causes the Mona Lisa effect [6] such that the user feels that he/she is seen from the image. Thus, C1, which derives the effect, is used to represent the behavior of looking at the user in S2 or S3. C1 is also used for the behavior of looking forward in S1. C2 and C3 are repeatedly used for the behavior of looking at the surroundings in S1. C2 or C3 is also used for the behavior of turning the face toward the user in S2. C4 is the behavior of bowing to the user in S3. Our method joins the video clips by smoothly switching them, and controls the speed of replaying the behaviors by adjusting to the movement of the user. The details of our method are described below.

3.2 Joining and controlling the video clips

We describe our method for joining the video clips of the image-based avatar. When switching from one video clip to another by simply connecting them, a discontinuous motion is caused because there is a gap between the initial postures in

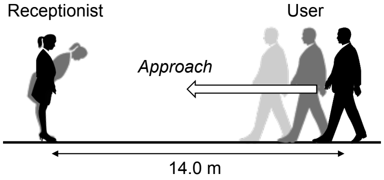

Fig. 3 Setup for measuring the parameters of our method.

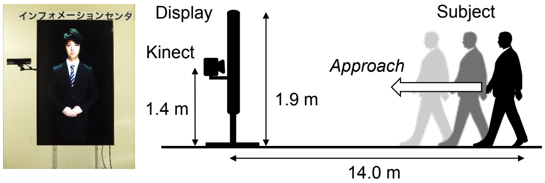

Fig. 4 Setup for the subjective assessment using the image-based avatar.

the video clips. Our method interpolates between the initial postures using an optical flow technique and noise reduction technique.

Next, we describe our method for controlling the speed of replaying the behaviors. To represent the behavior of looking at the surroundings in S1, our method repeats C2 and C3 within a cycle of T_1. To represent the behavior of turning the face toward the user in S2, our method fast forwards or rewinds the video clip that is currently playing. We finish this operation within $T_2 = (D_1 - D_2)/v$, where v is the user's walking speed. T_2 is automatically computed each time using v measured in real time using a depth sensor. To represent the behavior of bowing to the user in S3, we simply play C4 within T_3. To represent the remaining behaviors, we continuously play the video clip of C1. We determine T_1, T_3 by observing the length of time of the receptionist's behavior.

4. Experimental results

4.1 Measuring the parameters of our method to determine the thresholds

We measured the parameters of the behaviors at the beginning of the interaction to determine the thresholds of our method. We assumed that the receptionist guided the user in the information center. We collected two receptionists (average age 23.5 ± 0.5; one male and one female; Japanese) and five users (average age 23.5 ± 0.5; four males and one female; Japanese). We explained the situation to each user, that is, the user was to ask the receptionist for the location of the check-in counter at the airport. We assumed that the user was approaching from the front of the receptionist. Figure 3 shows the setup for the measurement of the parameters. The user started walking from a position 14 m away from the receptionist, approached the range for which he/she felt it was easy to speak, and spoke to the receptionist. We acquired the parameters of our method twice for each user. From the results of the measurements, we obtained $D_1 = 5.6 \pm 1.6$ m, $D_2 = 3.9 \pm 0.4$ m, $A = 20$ dB in 0.08 s, $T_1 = 2.0 \pm 0.3$ s, and $T_3 = 2.7 \pm 0.3$ s. Note that the average walking speed of the user was $v = 1.4 \pm 0.1$ m/s. We used the averages of the parameters for the thresholds of our method.

4.2 Experimental conditions for the subjective assessment using the image-based avatar

To conduct the subjective assessment of the beginning of the interaction, we collected 14 subjects (average age 23.0 ± 1.3; 12 males, and two females; Japanese) who did not participate in the experiment in Section 4.1. We explained the situation to each user, that is, the image-based avatar would guide him/her in the information center. Figure 4 shows the setup for a subjective assessment using the image-based avatar. We played the video sequences of the avatar on a vertically placed 80-inch display (Sharp PN-A601). To easily find the avatar from a distance, we attached the title 'information center' at the top of the display. To measure the distance from the subject to the avatar and the audio power of the subject, we used a depth sensor and microphone (Microsoft Kinect v2). Each subject started walking from a position 14 m away from the avatar, approached the avatar, and spoke to the avatar. We tested using male and female avatars for each subject.

We conducted the subjective assessment to compare the following image-based avatars:

- **I1**: Look forward continuously.
- **I2**: Look forward and bow to the user when the user approaches the avatar.
- **I3**: Look at the surroundings, turn the face toward the user, and look at the user when the user approaches the avatar.
- **I4**: Look at the surroundings, turn the face toward the user, and bow to the user when the user approaches the avatar.

I1 was a baseline method that did not embed the awareness state and response state. I2, I3, and I4 were methods that embedded both states, but the behavior in each state was different. Each subject repeated approaching and speaking to each avatar. Figure 5 shows the video sequences of I1 to I4. We used the video clips as C1: S1 to S3 in I1 and S3 in I3; C2 and C3: S1 and S2 in I3 and I4; C4: S3 in I2 and I4. Note that we simply displayed a message in S4 for all video sequences. After beginning of the interaction was complete, we asked the subjects the following questions:

- **Q1**: Was it easy to understand the timing when speaking to the image-based avatar?
- **Q2**: Did the avatar resemble the behavior of a real receptionist?
- **Q3**: Was it possible to smoothly start the interaction with the image-based avatar?
- **Q4**: Did you feel that the image-based avatar behaved politely when starting the interaction?

Each subject provided a rated score using five response levels (1: disagreeable, 6: agreeable) for each question. We also

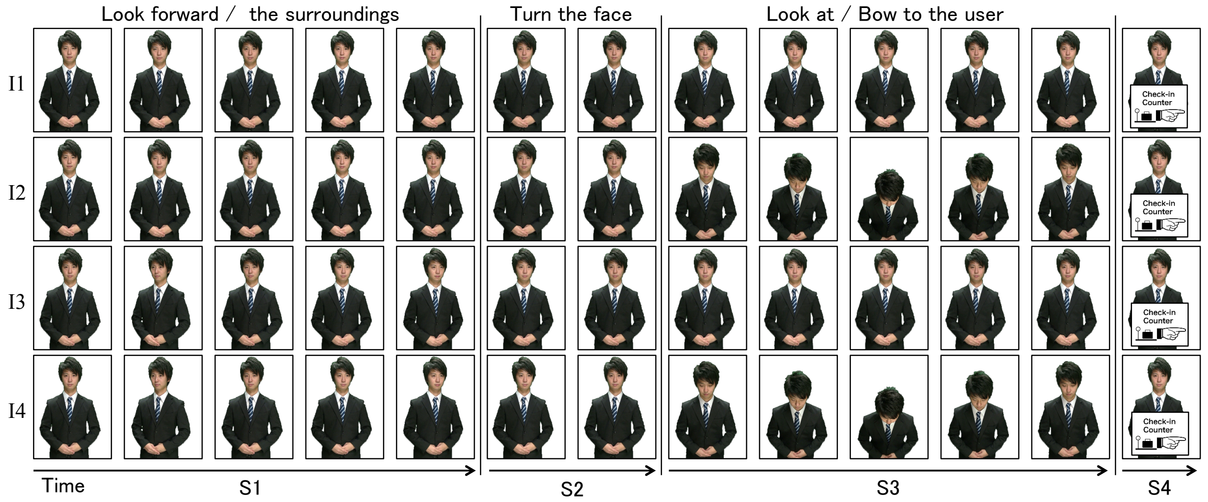

Fig. 5 Video sequences of the image based-avatar for the subjective assessment.

asked the inverse questions of Q1 to Q4. We randomly selected the order in which the video sequence of the avatar was played for each user.

4.3 Results of the subjective assessment using the image-based avatar

Figure 6 shows the average and standard deviation of the subjective scores for each question. The scores of the inverse questions were inverted and added to the corresponding scores for Q1 to Q4. We used the Friedman test, the Wilcoxon signed-rank test and Bonferroni correction after checking the normality of the scores. Comparing the behaviors of looking forward (I1) and looking at the surroundings (I3), marginal differences ($p < 0.05$) emerged for Q2 and Q4, but did not emerge for Q3. We believe that it is less effective for the image-based avatar to perform these behaviors in S1. Comparing the behaviors of bowing to the user (I2, I4) and looking at the user (I1, I3), significant differences ($p < 0.01$) emerged for all questions. Thus, bowing in S3 was more agreeable to the user. We believe that it is important to embed the awareness state (S2) and response state (S3) in the image-based avatar at the beginning of the interaction.

5. Conclusions

We proposed a method for embedding the awareness state and response state in an image-based avatar to smoothly start an interaction with a user. We designed a model of the receptionist's behaviors by observing an interaction in an information center. Our method joined and controlled video clips to replay the receptionist's behavior. Experimental results demonstrate that the use of the awareness state and response state worked well for the image-based avatar system. In future work, we will expand our assessment of the interactive system, and we intend to develop a method to combine the behaviors in the action state.

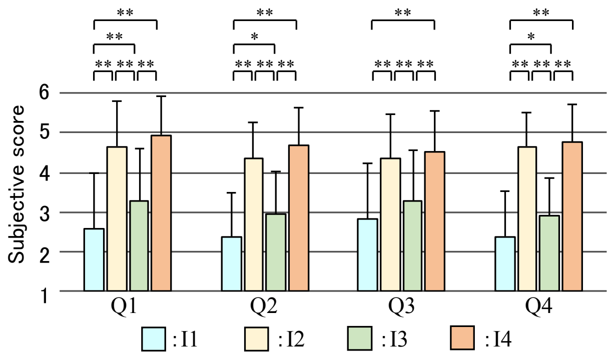

Fig. 6 Results of the subjective assessment ($** : p < 0.01, * : p < 0.05$) using the image-based avatar.

References

[1] S. Robinson, D. Traum, I. Ittycheriah, and J. Henderer, "What would you ask a conversational agent? observations of human-agent dialogues in a museum setting," Proceedings of the Sixth International Conference on Language Resources and Evaluation, pp.28–30, 2008.

[2] A. Lee, K. Oura, and K. Tokuda, "Mmdagent - a fully open-source toolkit for voice interaction systems," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.8382–8385, 2013.

[3] R. Artstein, D. Traum, O. Alexander, A. Leuski, A. Jones, K. Georgila, P. Debevec, W. Swartout, H. Maio, and S. Smith, "Time-offset interaction with a holocaust survivor," Proceedings of the 19th International Conference on Intelligent User Interfaces, pp.163–168, 2014.

[4] A. Jones, J. Unger, K. Nagano, J. Busch, X. Yu, H.I. Peng, O. Alexander, M. Bolas, and P. Debevec, "An automultiscopic projector array for interactive digital humans," ACM SIGGRAPH 2015 Emerging Technologies, p.6:1, 2015.

[5] M. Nishiyama, T. Miyauchi, H. Yoshimura, and Y. Iwai, "Synthesizing realistic image-based avatars by body sway analysis," Proceedings of the Fourth International Conference on Human Agent Interaction, pp.155–162, 2016.

[6] V. Bruce and A. Young, In the eye of the beholder: The science of face perception, Oxford University Press, 1998.