Hand Contact between Remote Users through Virtual Avatars

Jihye Oh, Youjin Lee, Yeonjoon Kim, Taeil Jin, Sukwon Lee and Sung-Hee Lee
Graduate School of Culture and Technology (GSCT)
Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
{ ojh1016, yjin1588, yeonjoon, jin219219, sukwonlee and sunghee.lee }@kaist.ac.kr

ABSTRACT
We present an avatar animation technique for a telepresence system that allows for the hand contact, especially handshaking, between remote users. The key idea is that, while the avatar follows the remote user's motion normally, it modifies the motion to create and maintain hand contact with the local user when the two users try to engage hand contact. To this end, we develop the support vector machine (SVM)-based classifiers to recognize the users' intention for contact interaction, and online motion generation method to create realistic image sequence of an avatar to realize the continuous contact with the user. A user study has been conducted to verify the effect of our method on the social telepresence.

CCS Concepts
• Computing methodologies~Motion processing • Computing methodologies~Virtual reality • *Human-centered computing~Virtual reality*

Keywords
3D Telepresence, Avatar interaction, Contact interaction, Character animation

1. INTRODUCTION
Teleconference or telepresence technology allows remote users to communicate with each other. Conventional telepresence systems, such as video conference system, enable users to watch other participants over the screen, but have limitations in that the users cannot interact over the screen, which severely limits the possible scope of interaction. Telepresence robots allow physical interaction but a sophisticated machine is necessary [4].

The virtual avatar has a high potential for enabling close interactions between remote users, and thus many researchers have developed avatar-based telepresence systems. However, there is a lack of research on implementing continuous contact interaction between remote users, not wearing haptic devices. In fact, it is extremely difficult for remote users to perform a proper contact interaction using only visual feedback. A representative case is handshaking: it is quite challenging to make a synchronized handshake movement only with a visual feedback due to the difference in speed and amplitude of the individual hand swing motion. The unsynchronized movement will decrease the sense of coexistence.

To solve this problem, we propose a novel avatar motion generation technique for social interaction. Our method allows the avatar to create an adaptive contact behavior online from the remote user's real-time motions. For this, two problems need to be solved. One is how to identify whether two people try to make contact interaction, and the other is, if so, how to create an appropriate motion of an avatar.

We address the first problem by developing classifiers to detect the people's intention to make or discontinue the contact interaction. SVM-based classifiers predict the intention given the captured motion of the users. To solve the second problem, we develop an avatar animation method using inverse kinematics to make the pose of the avatar's arm during the contact.

CASA 2016, May 23-25, 2016, Geneva, Switzerland
 ISBN 978-1-4503-4745-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2915926.2915947

2. RELATED WORK
There are a number of ways to interact with a remote user in telepresence system. A projection-based multi-user system may allow for the mutual understanding of pointing and tracing gestures by using depth and color cameras [1]. This system provides multiple users with prospectively correct stereoscopic images but does not enable contact interactions. A video-based telepresence system with a haptic robotic arm enables the users to shake hands with the remote user [5], but the immobile mechanical system usually lacks in expressiveness and limits the user's range of activity to the area in front of the system.

There has been research for real-time interaction between a human and a virtual avatar. In [2], a virtual avatar responds to the user-input motion requiring close interaction such as partner dancing and fighting. To coordinate the avatar with the user, the user's intention is predicted and the avatar's motion is retargeted. Similarly, in [6], virtual avatar's motion is created to interact with a user in real-time by retrieving suitable motion capture data of two interacting persons. Our work differs from the previous work in that the avatar motion is generated not from a motion database but by referring to the motions of both of the users.

3. METHOD
We describe the overall procedure for the offline and online processes, and then explain how we trained the SVM-classifiers to predict contact and separation of the hands.

3.1 System Overview
We have built a prototype telepresence environment using only commodity sensors and display devices. Both of the remote users

can see and interact with each other's avatar through the Oculus Rift DK2 and Kinect v2. In online process, distance features are extracted from the captured motions and analyzed with classifiers as shown in Figure 1. When the users are not in contact with each other, the contact classifier runs and predicts whether they try to make hand contact or not. If predicted to be negative, the raw motions of the users are directly applied to the avatars. If predicted to be positive, the opposite avatar' arm is adjusted to the local user's motion to initiate and maintain contact. When the users are in contact phase, the separation classifier runs and checks whether the users attempt to cease the contact, in which case the avatar returns back to follow its owner's motion.

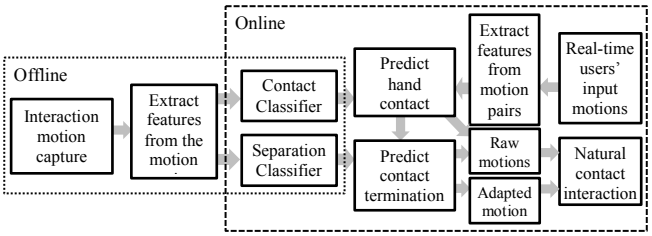


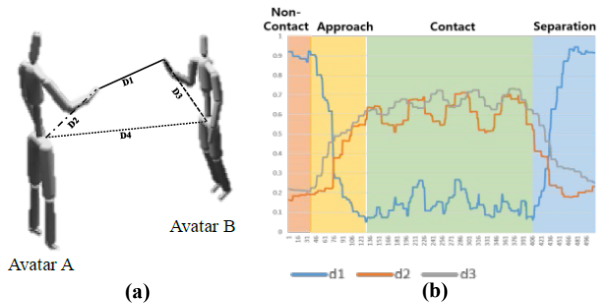
Figure 1. System overview.

3.2 Contact and Separation Classifiers

Given a temporal sequence of geometric features from the spatial relationship between avatars (or users), contact and separation classifiers predict the people's intention to make or discontinue the contact interaction.

3.2.1 Feature extraction and phase division

We collect various contact and non-contact motion pairs between two users to train the classifiers in the offline process. Training data includes handshake motion pairs as well as non-contact motions such as gestures. The features extracted from motion data consist of four distances as shown in Figure 2a. While the distance D1 between the two hands is an obvious feature for the hand interaction, we added the distance between the hip and the hand (D2, D3) to improve the performance of the classifiers. Eventually, we determine the features as d1, d2, and d3 by normalizing D1, D2, and D3 with root distance between the users (D4). These features show common pattern over time among handshake motion pairs as shown in Figure 2b.



(a) (b)

Figure 2. The distance features (2a) and temporal feature pattern during the handshaking (2b).

To train contact and separation classifiers, the temporal trajectories of features are divided into four phases: non-contact, approach, contact, and separation phases. The contact and non-contact phases refer to the states that the hands are steadily in contact or not. The approach and separation phases refer to the transition between the contact and non-contact phases. In real-time, the current phase is identified and the appropriate classifier is used. The contact classifier runs in the non-contact phase to detect the approach phase, and the separation classifier runs in the contact phase to recognize separation phase.

3.2.2 Feature vector labelling

To consider the sequence of the features, we form a feature vector composed of features (d1, d2, d3) for 15 frames, and thus the dimension of a feature vector is 45. The feature vector is updated every frame using a sliding window technique, i.e., a feature vector at a certain frame is constructed by removing the features of the oldest time step from the previous feature vector and adds the latest data into the vector, thereby keeping the total length of the window constant (Figure 3a). Every feature vector is labelled as the majority among the phases of the 15 frames. For example, the feature vector labelled with 2 (contact state) would change into 3 (separation state) when more than 8 frames of all 15 frames are in the separation phase (Figure 3b).

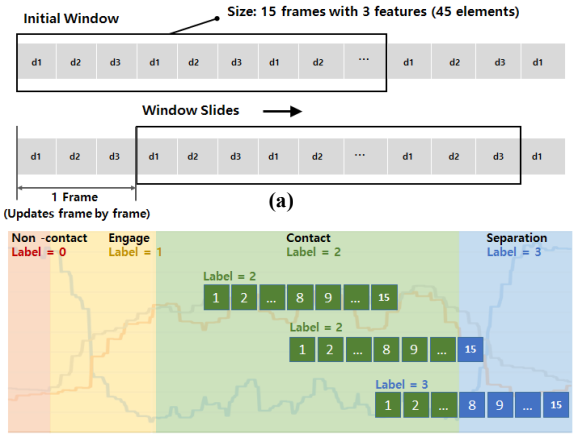


Figure 3. Sliding window of features (3a) and example labellings of feature vectors (3b).

3.2.3 Supervised Learning of Classifiers

The SVM is used to train the contact and separation classifiers. We trained contact and separation classifiers with 5006 and 3806 observations, respectively. The specific numbers of training data and test data are shown in Table 1. The data is manually labelled into each phase according to the features' pattern.

Table 1. The number of data used for training and testing the classifiers.

Contact Classifier	Training Data	Test Data
Non-Contact	934	208
Approach	4072	160

Separation Classifier	Training Data	Test Data
Contact	2388	526
Separation	1418	291

The contact and separation classifiers are also trained with other methods: K-nearest neighbor, decision tree, and naïve Bayes classifiers. Among these learning models, SVM-classifier shows the highest prediction rates (92.55% for contact and 99.24% for separation classifier) as shown in Table 2.

Table 2. Accuracy test of four learning models.

Classifier	SVM	KNN	Decision Tree	Naïve Bayes
Contact	0.9255	0.4362	0.4468	0.7766
Separation	0.9924	0.9542	0.9466	0.3690

3.3 Generating Contact Motion of Avatars

We perform inverse kinematics to generate the remote avatar's arm pose during the approach and contact phases. Inverse kinematics is carried out by minimizing the following cost function f

$$f = w_t\|\mathbf{p}(\boldsymbol{\theta}) - \mathbf{p}^*\|^2 + w_h\|\mathbf{u}(\boldsymbol{\theta}) - \mathbf{u}^*\|^2 + w_p\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$$

where $\boldsymbol{\theta}$ denotes the joint angle vector of the arm. The first term on right hand side drives the hand position $\mathbf{p}(\boldsymbol{\theta}) \in \Re^3$ to reach the target position $\mathbf{p}^*$ to naturally touch the partner's hand. Next two terms ensure naturalness of the poses; the second term drives the avatar's upper arm direction $\mathbf{u}(\boldsymbol{\theta}) \in \Re^3$ to follow the partner's upper arm direction $\mathbf{u}^*$. The third term is a regularizer to encourage the joint angles $\boldsymbol{\theta}$ to be the default value $\boldsymbol{\theta}^*$, which are set to zeros in our experiment. The weights w_t, w_h and w_p control the significance of each term. We solve the optimization problem at each time frame using the Levenberg-Marquadt algorithm.

4. EXPERIMENT

We predicted that the motion adaptation would have a positive effect on social telepresence. We conducted an experiment to verify the effect of our avatar animation technique to enhance social telepresence. Specifically, we considered three aspects of social telepresence: feeling of coexistence, plausible motion, and intimacy.

4.1 Questionnaire

The questionnaire asks the extent to which the subject's impression corresponds with the statement. The three aspects of social telepresence are considered in measuring the degree of telepresence. The questionnaire includes the following statements.

- I felt as if I were facing the presenter in the same room [3].
- I felt as if I were shaking hands with the remote user in real life [4].
- I felt as if I were becoming close to the remote user in the same room [4].

We also added the following statements for prior information to check how much the subjects are used to the equipment used in the experiment and to the virtual environment.

- I am familiar with the virtual reality equipment such as HMD.
- I could be fully aware of the presenter's action.
- I could take action in virtual reality as I wanted.

All the statements were rated on a 7-point Likert scale where 1 = strongly disagree, 4 = neutral, and 7 = strongly agree.

4.2 Subjects

The experiment was a within-subject design. All participants tested both conditions, but half started with the control condition and the other half started with the adapted motion condition. The subjects were twenty-four graduate students whose ages ranged from twenty-two to thirty-three.

5. RESULTS

We compared the results of the control condition and the adapted motion condition as shown in Figure 4. When the user tries to make a continuous handshake in the control condition, the gap between two avatars' hands is easily observed (Figure 4a). With our techniques, the user can easily make a proper hand contact pose because the partner's avatar adjust its pose to create and maintain contact (Figure 4b).

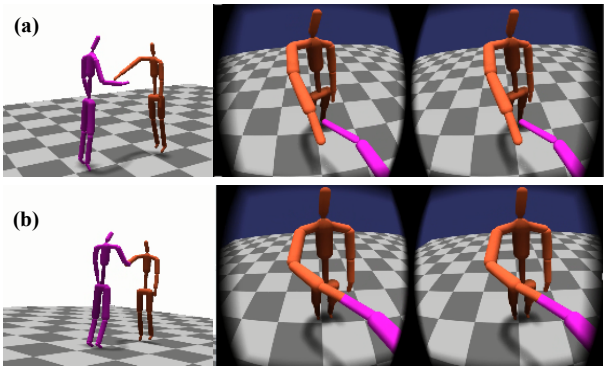


Figure 4. Screenshots of the control condition (4a) and the adapted motion condition (4b) in handshaking. (Left: screen view, Right: HMD view)

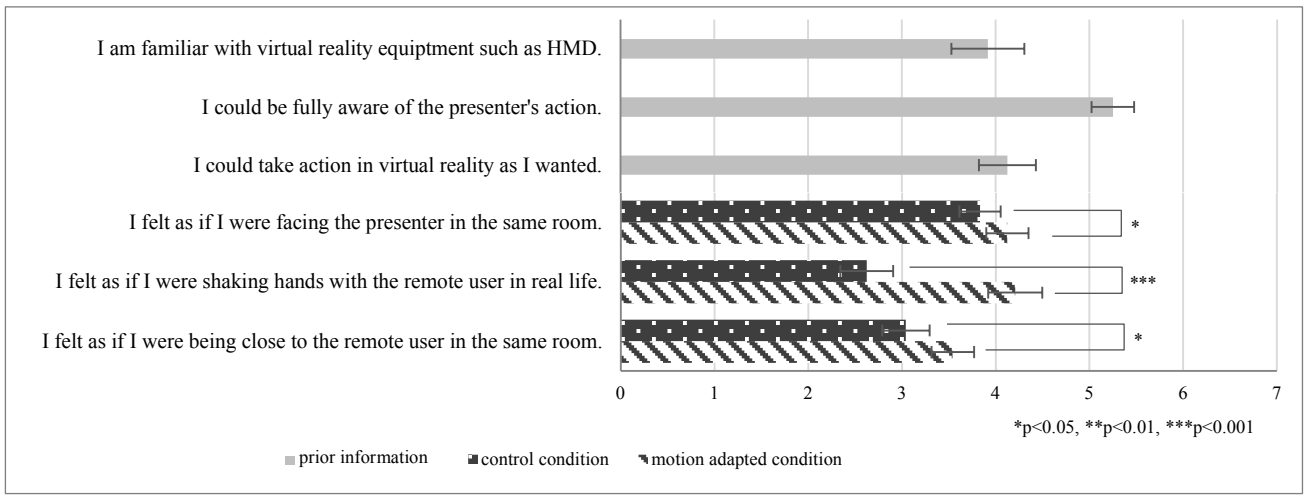


Figure 5. Results of the questionnaire (motion adaptation factor).

5.1 Results of the User Study

We used one-way within-subject ANOVA to verify the effect of motion adaptation technique to the subjects' immersion in the telepresence. The results of the questionnaire are shown in Figure 5 in which each box represents the mean value of the responses to each statement, and each bar means the standard error of the mean value. We set within-subject factor as motion adaptation, and the significance threshold was set to 0.05.

First, judging from the preliminary survey, the subjects were not expert at handling the device such as a HMD but their familiarities with the device were up to par. Similarly, they were able to have some control of their own behavior in a virtual environment. In particular, the action recognition of the other presenter was relatively higher. For the overall subjects, they were not to the extent of experts for the equipment and the virtual environment, but were accustomed to using them to a certain level.

Second, the overall scores for feeling of coexistence, plausible motion and intimacy were lower than 4 (neutral) in control condition. Some subjects commented that "The skeleton basically takes away from a sense of closeness" and "The face of someone that I know will bring a better response." When applying the motion adaptation, the largest increased aspect of telepresence was plausible motion. We found that our motion adaptation technique had effect on the feeling of coexistence ($F(1, 23)=6.749$, $p=0.016$), plausible motion ($F(1, 23)=22.380$, $p<0.001$), and intimacy ($F(1, 23)=5.308$, $p=0.031$). The subjects' additional statements were that "When I saw the opponent's behavior in response to my movement, I felt that there is the other person in the other side of the screen" and "I felt a sense of intimacy to see the opponent react even when I took the action of the non-handshake such as waving a salute." The results ultimately ended up in near 4 (neutral) but, as we hypothesized, motion adaptation has positive effects on social telepresence.

6. CONCLUSION

Our research achieves the social telepresence system in which remote users can seamlessly shake hands by predicting hand contact and adapting avatar motion. Our techniques have some limitations. Currently, the hand contact interaction is limited to a handshake: we plan to extend the range of available hand contact, such as high-five. The accuracy of the classifiers also needs to be improved so that it can distinguish a handshake from similar-looking motions such as the motion of rock-paper-scissors. As a future work, finger-to-finger interaction between the remote users' avatars will be considered for more natural interaction, and the actual appearance of the opposite user will be displayed as for the avatar's appearance instead of the stick figure. Despite these limitations, we believe that our method expands the possibilities of the avatar-mediated social telepresence by allowing for the hand contact between the remote users.

ACKNOWLEDGMENTS

This work was supported by the Global Frontier R&D Program funded by NRF, MSIP, Korea (2015M3A6A3073743).

REFERENCES

[1] Beck, S., Kunert, A., Kulik, A., and Froehlich, B. 2013. Immersive group-to-group telepresence. *Visualization and Computer Graphics, IEEE Transactions on*. 19, 4, 616-625.

[2] Ho, E. S., Chan, J. C., Komura, T., and Leung, H. 2013. Interactive partner control in close interactions for real-time applications. *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM) (June. 2013). 9, 3:21. DOI= http://dl.acm.org/citation.cfm?id=2487274.

[3] Nakanishi, H., Kato, K., and Ishiguro, H. 2011. Zoom cameras and movable displays enhance social telepresence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, May 07 – 12, 2011). CHI '11. ACM, New York, NY, 63-72. DOI= http://dl.acm.org/citation.cfm?id=1978953.

[4] Nakanishi, H., Tanaka, K., and Wada, Y. 2014. Remote handshaking: touch enhances video-mediated social telepresence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Canada, April 26 – May 01, 2014). CHI '14. ACM, New York, NY, 2143-2152. DOI= http://dl.acm.org/citation.cfm?id=2557169.

[5] Tanaka, K., Nakanishi, H., and Ishiguro, H. 2014. Robot conferencing: physically embodied motions enhance social telepresence. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (Toronto, Canada, April 26 – May 01, 2014). CHI EA '14. ACM, New York, NY, 1591-1596. DOI= http://dl.acm.org/citation.cfm?id=2581162.

[6] Vogt, D., Grehl, S., Berger, E., Amor, H. B., and Jung, B. 2014. A data-driven method for real-time character animation in human-agent interaction. In *Intelligent Virtual Agents* (August. 2014). Springer International Publishing, 463-476. DOI=http://link.springer.com/chapter/10.1007/978-3-319-09767-1_57.