Control of a Telepresence Robot Using Force data

Takuya Yokoyama [*] **Vincent Hernandez** [*,**] **Liz Rincon** [*] **Gentiane Venture** [*]

[*] *Tokyo University of Agriculture and Technology, Japan (e-mail: venture@cc.tuat.ac.jp)*
[**] *Surfclean, Japan*

Abstract:
Telepresence robots are robots intended to compensate for non-verbal information during telecommunication. However, current telepresence robots don't have sufficient functionality to send gesture information, within non-verbal information. This research aims to develop a communication system that recognizes the motion of the human and supplements the lack of gesture information by transmitting it to humanoid robots. The method proposed involves motion data acquisition using force data, gesture recognition with CNN (Convolutional Neural Network) and control of a humanoid robot with the transmission of gesture by on-line control. Finally, the proposal is evaluated by the TDMS (Two-Dimensional Mood Scale) to verify the difference from using the current telepresence robot. As a result, we recognized 6 motions with an automatic motion recognition accuracy of 77.8%. Telepresence using a humanoid robot was confirmed to improve comfortable feeling by transmitting a gesture, although a significant difference from existing telepresence robot was not confirmed.

Keywords: Telecommunication, Robot control, Force, Neural networks, Real-time AI

1. INTRODUCTION

The declining birthrate and aging population are a concern regarding the decline of the working population in most OECD countries Alfredsson and Winther (2019). As part of the counter measures, telework, which uses ICT (Information Communication Technology) and is a flexible way of working was recommended. Telework Beauregard et al. (2019) not only uses time and space efficiently, it also minimizes transport being also good for the environment to a certain extent. Despite all the benefits of telework, its spread has been stagnant at present, and the reinforcement of information system is required. One alternative is the telepresence technology, which gives the sense of presence as if you are directly facing a remote person. It has the role of compensating for the lack of non-verbal information and a sense of being present lacking in telephone and image communication. One of the examples of its use is the telepresence robot (Herring, 2013) that consists in a communication between a person in a remote location with the possibility to operate the robot and interact with a local person. Several telepresence robots, such as Double, Beam Enhanced, PadBot, Supercam... have started to hit the market but with moderate success. Indeed, existing telepresence robots are usually a screen on a stick with some mobility abilities. They do not have parts equivalent to human arms and neck, and are not capable of making gestures, which are an important element in dialogues and verbal communication Björnfot and Kaptelinin (2017).

We believe that this problem could be solved if we use robots that moderately resemble humans and that can easily transmit gesture information in synchronization with human movement. To transmit human motions it could be teleoperated intentionally by the human or the human movements could be measured and transmitted to the robot directly. In the latter case most human mimicking systems use kinematics information, but we propose to use force information. The goal of this paper is to use a partially humanoid robot to compensate for the missing gesture information of conventional telepresence robots without arms and to realize more comfortable communication, in particular during given a presentation or a lecture. To transmit motion from the human to the robot as fluently and intuitively as possible we propose to use force information with the ground. The motion is measured by a force sensor, e.g. Nintendo Wii Balance Board. The collected data are segmented and recognized with a deep learning CNN (Convolutional Neural Network) LeCun et al. (2015) and set to control the robot. We finally tested our proposed method on a small sample population and compared the results obtained with a conventional stick-screen telepresence robot and a robot that doesn't move.

2. RELATED WORKS AND CONTRIBUTION

2.1 On non-verbal communication

Birdwhistell (2010) analyzes that only 35% of the information is verbally transmitted in the dialogue between individuals, and the remaining 65% is conveyed by nonverbal communication. Kret et al. (2011) showed that from a psychological point of view, body movement is as important as facial expression in communication using emotional expression. Hasegawa and Nakauchi (2014) operated the robot with two methods. One is using a controller and the other with a human body movement. So the first method can use only conscious movement, and the second method can use both conscious and unconscious (intuitive) movement. This work showed the importance to sent the information unconsciously by the body movement.

2.2 On motion recognition

In most work motion recognition is achieved using kinematics data such as motion capture data Ott et al. (2008), image data Chen et al. (2018) or IMU data Mascret et al. (2018). However, image needs a large area for measurement, is prone to occlusion, and needs better processing computers. The IMU cannot measure the whole body with a small number of devices, so if we want to increase the target motion, it is necessary to increase the number of sensors. Neither can be said to be really suitable for telepresence. On the other hand, motion recognition for simple tasks using force data has been verified by Yabuki et al. They proposed a segmentation method using dispersion, and Principal Component Analysis for the recognition of seven gymnastics motions with a difference in motion (Yabuki and Venture, 2015). We propose to use motion force similarly for this study. Though we are using a standing participant in our experiment, a force plate positioned under the buttock on a chair could similarly collect data from a seating participant.

2.3 On motion retargeting

Animating robots with human-like motions is often solved as a retargeting problem. In their review paper, Kulić et al. (2016) showed various approaches for motion retargeting. However, the kinematic approach is particularly difficult under disturbances, and furthermore, the destination robot cannot always follow completely the human in the physical environment due to inherent differences in the body structure of the robot and the human. Therefore retargeting may not be adequate when the robot must be expressive as no retargeting method guarantees the expressivity of the movement to our knowledge. In this preliminary study we are using a set of predefined motions, while we will focus in the future in the proper retargeting.

2.4 On user's mood evaluation

Mood and emotion evaluation has been the focus of several research. One of the most popularly used methods is the SAM (Self-Assessment Manikin) Morris (1995) that allows subjects to directly respond about their degree of pleasure, arousal, and dominance. Another approach is the TDMS (Two-Dimensional Mood Scale) Sakairi et al. (2013) that was implemented to evaluate the change of subject's mood and applied during experimental evaluation. By intuitively answering about eight items, the TDMS can measure the pleasure, arousal, vitality, and stability of the subject's mood. The degree of pleasure and arousal can be calculated from the degree of vitality and stability and be plotted on a two-dimensional coordinates as shown in Figure 1. It is often used to see a change before and after stimulation. In this study, we therefore use the TDMS test.

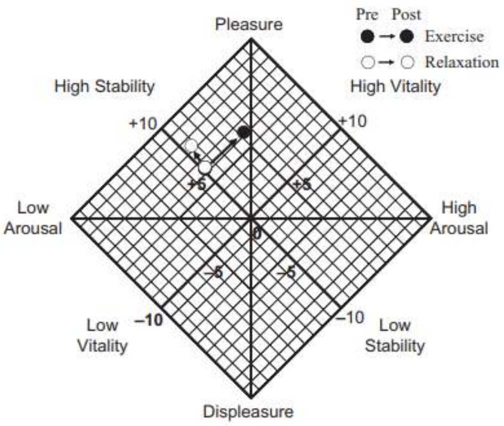


Fig. 1. TDMS example: Changes in mood states with exercise and relaxation. Sakairi et al. (2013)

2.5 On deep learning

Deep learning is often used especially for image recognition LeCun et al. (2015). In particular, image recognition in real-time is performed using SSD (Single Shot Multi Box Detector) Liu et al. (2016) or YOLO (You Look Only Once) Redmon et al. (2016). They perform segmentation and recognition of objects at one time, but in this research, we decided to sequentially perform segmentation and recognition on force data instead of image data using a CNN that is simpler and easier to adjust the algorithm.

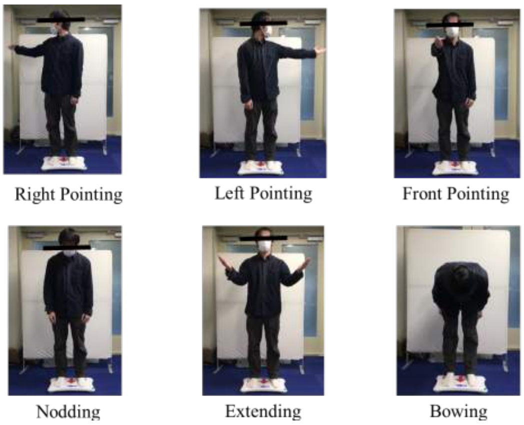
2.6 Contribution

As mentioned above, our purpose is to prove the usefulness of using a robot that conveys human gestures for telepresence by demonstrating an example and evaluating the receiver's mood. In this research, in order to add nonverbal information to a telepresence robot, we tried to synchronize the movement of the human and the humanoid robot. The robot's motion is generated in advance, and is called by the recognized motion class, so there are no kinematic problems. Motion recognition is performed by classifying force data into CNN. We also conducted a user study using the TDMS test to evaluate the perception of some users.

3. AUTO SEGMENTATION AND MOTION RECOGNITION

3.1 Methods

For segmentation and recognition of the human motions, two CNNs were used to predict behavior from the force data acquired in real-time. This time, we used a popular and easy-to-understand CNN as an introduction, but in the future, we will try our proposal using recurrent models suitable for time series data. CNN models were based on VGG16 Simonyan and Zisserman (2014). We manually tuned some parameters that shows in the next section. These parameters were adjusted to get the recognition rates over 80%, which is similar to the past research Yabuki and Venture (2015). In this case, there were few recognition actions, but it is considered that the optimization by grid search is required if the number



Right Pointing
Left Pointing
Front Pointing
Nodding
Extending
Bowing

Fig. 2. Gestures for the human motion recognition using force data

of actions increases as recognition becomes difficult. The total force and the Center of Pressure (CoP) were collected for learning data using a Nintendo Wii Balance Board, which is an inexpensive force sensor device. This sensor's sampling rate is 30 fps. Five participants, from 22 to 23 years of age (average age: 22. o.), all males, took part in the study for data collection. This data is used for the development of the learning models. All of the participants were students at Tokyo University of Agriculture and Technology. Of the acquired data, 80% was used for training and 20% for testing. We used training data for hyperparameter optimization. The amount of data was 4800 samples for the segmentation model and 1200 samples for the recognition model.

3.2 Application

In this research, six motions were chosen for the motion recognition as inspired by Tian and Bourguet (2016). They are shown in Figure 2. The motions were *Pointing* (using three types) and *Extending*, which were particularly frequent gestures when giving a lecture as Tian et al. suggested. In addition, to demonstrate that this system can be applied to other motions than the arms, we added *Nodding* and *Bowing* motion, the former using the neck joint and often used in conversation, the latter particularly useful in the Japanese context. All movements were limited to the upper body, and we instructed participants to avoid moving their feet. There was no specification of the arm/head/torso angle or speed at which they should achieve the movements. First, the degree of deviation $D[-]$ of the acquired force data from the subject's weight was obtained from the following equation for the manual segmentation.

$$D[-] = \left(1 - \frac{ForceData}{BodyWeight}\right)^2 \tag{1}$$

As Figure 3 shows, the start and end of the movement were first manually recorded from the waveform of the degree of deviation (Eq. 1). For this manual segmentation, the start and end points are detected by inspection when the data force is changing. The manual segmentation is used for the training of the segmentation algorithm and for the performance evaluation during testing as the ground truth. For the auto segmentation, a sliding window is created with the arbitrarily chosen size of 20 frames (0.6 sec) and it is moved each time by 1 frame. During the window motion, the start and final points that contain the data are detected. The auto-segmented motion data were used for the subsequent motion recognition training. Both of the learning used total force and CoP data. For auto segmentation's CNN the parameters were tuned as:

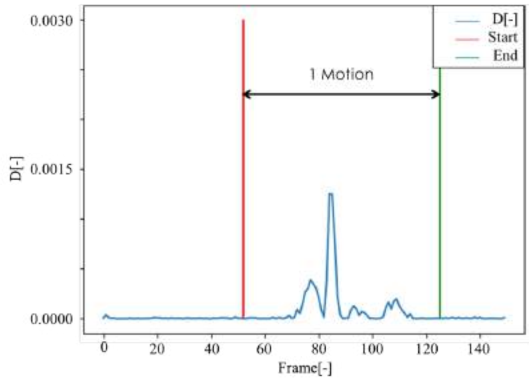


Fig. 3. Data segmentation for training motion models: Manually detected segment points in the change of the deviation of force data (Right Pointing motion)

- Number of convolution layers [7, 10, 13],
- Batch size [32, 64, 128],
- Convolution output dimension [32, 64],
- Dense layer output dimension [512, 1024, 2048],
- Dropout [0.4, 0.5, 0.6, 0.7]

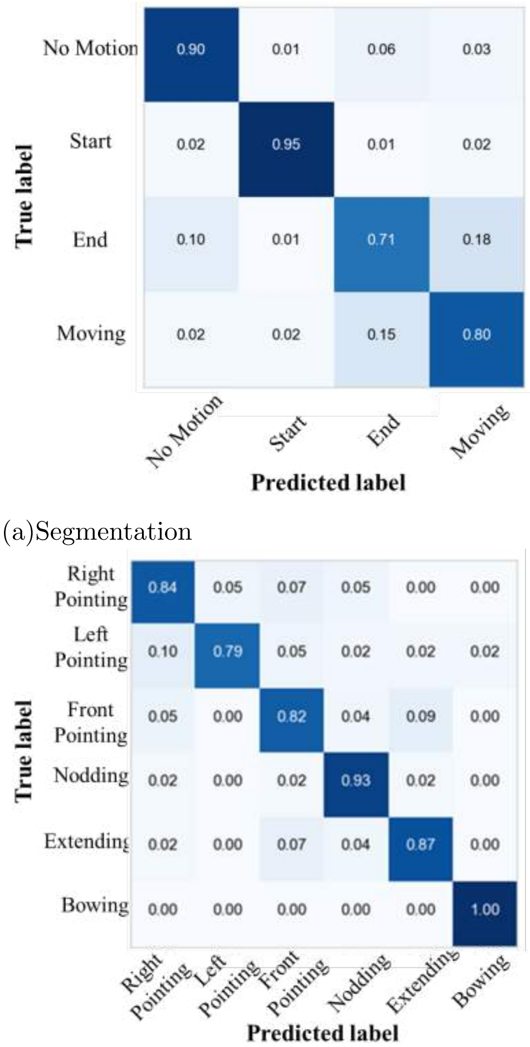
For motion recognition's CNN the parameters were:

- Number of convolution layer [3, 7, 10, 13],
- Batch size [16, 32, 64],
- Convolution output dimension [32, 64],
- Dense layer output dimension [36, 72, 144],
- Dropout [0.4, 0.5, 0.6, 0.7]

The kernel size is [1, 3], and the step is [1, 1] for both models.

Figure 4 shows the learning results. Segmentation's CNN used four classes: No motion, Start, End, and Moving. Testing data F-score was 0.83. Motion Recognition's CNN classified the 6 motions shown in Figure 2. Verification data F-score was 0.87.

We connected these two CNN in an off-line scheme. In auto-segmentation, data of the same size as the input at the time of learning was input to the CNN model for segmentation while shifting one frame at a time and output the probability of each class. Then, when the probability of Start among outputs exceeds an arbitrarily chosen value of 90%, this is taken as the start point. After that and similarly, the point where the probability of End exceeds 90% (arbitrarily chosen) is taken as the end point. We tested with the 20% unlearned data of the subject. The success rate for segmentation was 83.5% on average. And as a result of input the data which succeeded in auto-segmentation to the CNN model for Motion recognition, it is represented in Figure 5 and the accuracy became 77.8%. Our off-line scheme was just to test the performance of the algorithm. An on-line implementation of the same



(a)Segmentation

(b)MotionRecognition

Fig. 4. CNN results: (a) Confusion matrix for the Segmentation, (b) Motion recognition results using manually segmented data

algorithm therefore is expected to have a similar rate of accuracy and have similar performances.

4. ROBOT CONTROL IN REAL-TIME

The segmented and recognized motion should be sent to the robot for it to take action. In this research, we use Pepper (Pandey and Gelin, 2018) as a humanoid robot that is capable of reproducing, to a certain extent, gestures. And as shown in Figure 6, on-line communication was made by socket connection. The main systems in the structure are configured by: i) human data collection, ii) auto motion segmentation & recognition, and iii) robot control. The flow starts with the human data collection in C# used for the force measurement with the Wii Balance Board. The auto motion recognition with CNN (segmentation and recognition of the motions) is developed with Python 3. Finally, the robot is commanded using the SDK Python 2. In the SDK, at the moment, the movements of the robot in advance for each gesture as shown in Figure 7. The data is collected every 0.03 sec. So, the segmentation network output modifies the probability of each class every 0.03 sec. The motion recognition network output sets the 6 gesture's probability immediately after the segmentation network recognizes the endpoint. The motion delay almost equals to each motion's length. The minimum delay is 0.3 sec (nodding), and maximum delay is 3 sec (bowing).

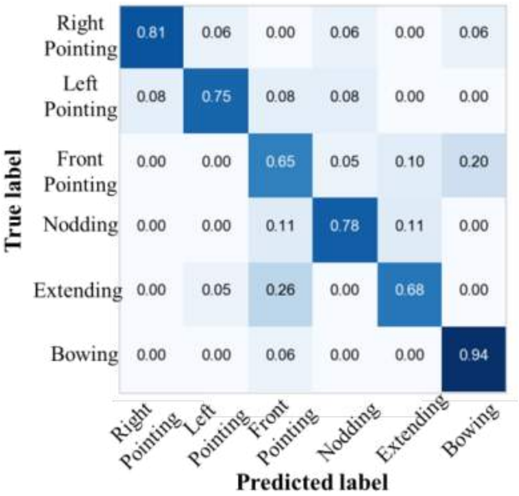


Fig. 5. Confusion matrix for the off-line auto segmentation and motion recognition result combined

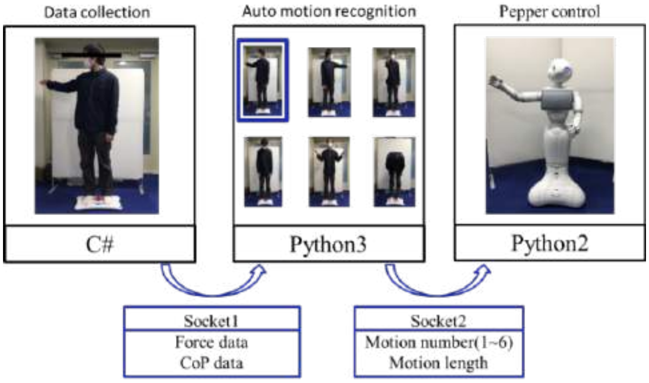


Fig. 6. On-line control structure to command the telepresence robot using the human motion recognition based on the measurement of force data

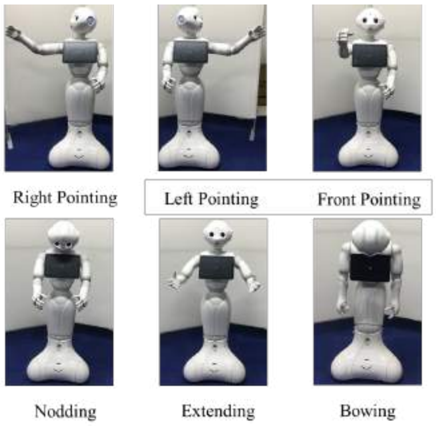


Fig. 7. Pepper robot movements corresponding to each of the 6 gestures

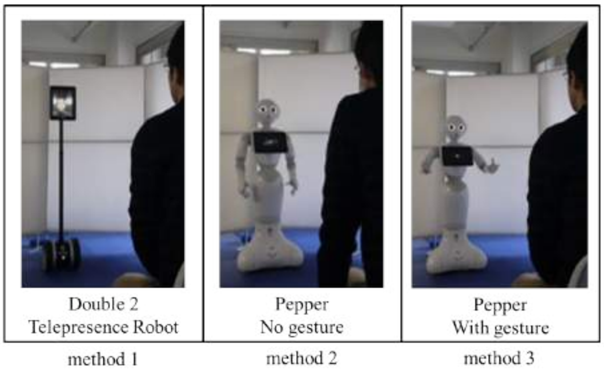


Fig. 8. Evaluation of different methods for the telepresence robot

5. EVALUATION EXPERIMENT

We conducted a user study to evaluate our proposal by comparing three telecommunication methods. As shown in Figure 8, the experiment was performed to measure and compare the changes in the mood of the receiving user before and after communication. The mood was measured using the TDMS test. The first telecommunication method uses the Double Robotics telepresence robot Double 2 and sends voice and image information by a small screen, which can send face and only their neck gesture information (hereafter "method 1"). The second telecommunication method uses Softbank humanoid robot Pepper without gesture transmission and facial information, only sending voice information ("method 2"). And the third telecommunication method is developed with the transmission of the human gestures to Pepper ("method 3"). Pepper has an 'Autonomous mode' which Pepper can move his arm slightly as a breathing human. We used it in the "method 2". So, in this case Pepper robot looks not inert.

For the validation of the proposal, twelve participants, from 20 to 26 years of age (average age: 22.6 y. o.), 11 males and 1 female, took part in the study. All of the participants were students at Tokyo University of Agriculture and Technology and gave their written consent. The same person delivered the talk as a remote user, and the subjects heard the story using the three methods, one subject at one time. It consists in reading out a specified text while performing the predetermined six actions at specified timing, and in performing a short free dialogue. All audio used the same external speaker. Participants experimented the 3 methods in a random order. All methods were performed with a 3-minutes break. They filled the TDMS questionnaire before and after of each experiment to calculate the change amount. The TDMS was conducted 6 times for each person.

6. RESULTS

As a result of one-way Anova and Tukey test, no significant difference could be identified in the change in the Arousal score as shown in Figure 9. However, comparing changes in the Pleasure score as shown in Figure 10, there are significant differences between the communication using either the telepresence robot "method 1" or the humanoid robot that performs gestures with our method "method 3" compared to the robot without sending gesture information "method 2". Although we found no significant difference between 'method 1" and "method 3".

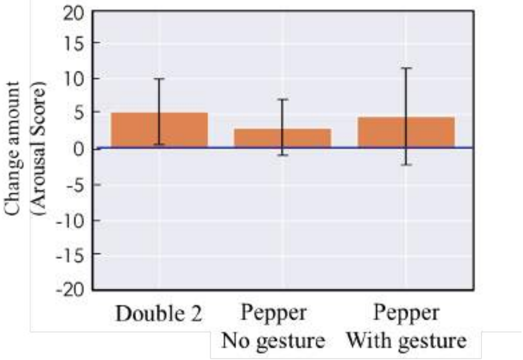


Fig. 9. TDMS result (Arousal)

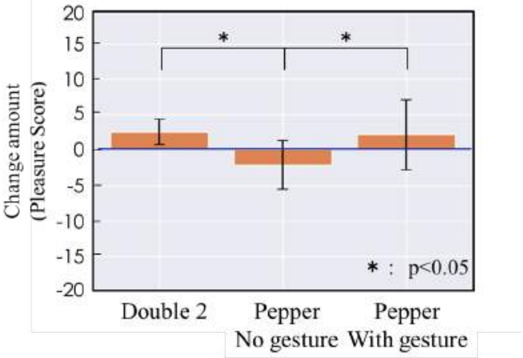


Fig. 10. TDMS result (Pleasure)

6.1 Discussion

From these results, it was confirmed that the receiving user had a tendency to feel as comfortable with our method than with a classic telepresence robot. The reason for the little difference may be that our robot doesn't show the face of the speaker and therefore if it can show body language it cannot show facial expressions, while the classic telepresence robot, while it doesn't have gestures it has facial expressions. Both seem to be equally important in the delivery of the verbal content. Would a robot that could do both certainly lead to even better results is something that needs to be clarified. Yet our work shows that transmitting even limited body gestures is a promising technique to improve the quality of interaction with telepresence robot as we originally assumed. Adding faces or facial expressions using the robot LED could also be useful and needs to be further investigated.

It can be seen from Figure 10 that the standard deviation is large particularly in "method 3". This indicates that some cases the pleasure was much higher when our algorithm performed well and synchronously with speech. However, the level of pleasure dropped significantly when sufficient gesture information was not transmitted. This was the case when motion recognition failed or when there was a large delay in the motion execution leading to a mismatch between body language and speech. This is a very promising result again for our method.

After the experiment, in order to know where to fix and why the variance is high, we asked participants to answer a few questions: (1) what they noticed about each method?, and (2) what can they mention about their concerns? As a result, regarding “method 3”, there were often concerns about the gap between the voice and the motions. As a countermeasure, it is conceivable to divide the specific motions into several parts, and increase the number of the split nodes as well as the start and the end points during the auto recognition. In addition, the delay reduction will be possible by increasing the timing at which Pepper robot starts the tracking. Others participants pointed out that the number of types of prescribed motions were small. A dynamic based motion retargeting is now under consideration in order to expand the number of motions.

7. CONCLUSION

Telepresence using a humanoid robot whose movements are controlled based on force data was successfully implemented with deep learning structures, and it was validated by the TDMS. Our real-time implementation for the auto segmentation reaches a performance of 83.5% success when performing specified gestures on the WiiBB. And the correct gesture was recognized with a success rate of 77.3%. Our user study measured the variation of the receiver’s mood at the time of telecommunication using TDMS with different telepresence methods. Our proposed method using the humanoid robot is comparable to the classic telepresence robot Double 2. However, there is a tendency to make the receiver feel more comfortable when motion information was successfully and timely transmitted to the robot. Therefore, although the proposed method of this research needs further improvements to overcome the existing limitations, it has proven very promising as a new mean of telepresence, in particular when delivering a lecture or a presentation and conversing casually. In the future, we aim to generalize the model by increasing the data and more particularly in using the dynamics information to generate the robot movements in a more systematic manner, so that the time delay will be removed and fluent communication can be achieved.

REFERENCES

Alfredsson, J. and Winther, A. (2019). Population ageing and average retirement age: A cross-sectional analysis on oecd countries.

Beauregard, T.A., Basile, K.A., and Canonico, E. (2019). Telework: outcomes and facilitators for employees.

Birdwhistell, R.L. (2010). *Kinesics and context: Essays on body motion communication.* University of Pennsylvania press.

Björnfot, P. and Kaptelinin, V. (2017). Probing the design space of a telepresence robot gesture arm with low fidelity prototypes. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, 352–360. IEEE.

Chen, J., Wang, G., Hu, X., and Shen, J. (2018). Lower-body control of humanoid robot nao via kinect. *Multimedia Tools and Applications*, 77(9), 10883–10898.

Hasegawa, K. and Nakauchi, Y. (2014). Preliminary evaluation of a telepresence robot conveying pre-motions for avoiding speech collisions. In *Proceedings of the Second International Conference on Human-Agent Interaction, Tsukuba, Japan*, 29–31.

Herring, S.C. (2013). Telepresence robots for academics. *Proceedings of the American Society for Information Science and Technology*, 50(1), 1–4.

Kret, M., Pichon, S., Grèzes, J., and de Gelder, B. (2011). Similarities and differences in perceiving threat from dynamic faces and bodies. an fmri study. *Neuroimage*, 54(2), 1755–1762.

Kulić, D., Venture, G., Yamane, K., Demircan, E., Mizuuchi, I., and Mombaur, K. (2016). Anthropomorphic movement analysis and synthesis: a survey of methods and applications. *IEEE Transactions on Robotics*, 32(4), 776–795.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., and Berg, A.C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Mascret, Q., Bielmann, M., Fall, C.L., Bouyer, L.J., and Gosselin, B. (2018). Real-time human physical activity recognition with low latency prediction feedback using raw imu data. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 239–242. IEEE.

Morris, J.D. (1995). Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research*, 35(6), 63–68.

Ott, C., Lee, D., and Nakamura, Y. (2008). Motion capture based human motion recognition and imitation by direct marker control. In *Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots*, 399–405. IEEE.

Pandey, A.K. and Gelin, R. (2018). A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics Automation Magazine*, 25(3), 40–48. doi:10.1109/MRA.2018.2833157.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Sakairi, Y., Nakatsuka, K., and Shimizu, T. (2013). Development of the two-dimensional mood scale for self-monitoring and self-regulation of momentary mood states. *Japanese Psychological Research*, 55(4), 338–349.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tian, Y. and Bourguet, M.L. (2016). Lecturers’ hand gestures as clues to detect pedagogical significance in video lectures. In *Proceedings of the European Conference on Cognitive Ergonomics*, 2. ACM.

Yabuki, T. and Venture, G. (2015). Human motion classification and recognition using wholebody contact force. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4251–4256. IEEE.