Sensors and Materials, Vol. 32, No. # (2020) 1–##
MYU Tokyo

S & M ####

Detection of Head Motion from Facial Feature Points Using Deep Learning for Tele-operation of Robot

Masahiko Minamoto,[1*] Shigeki Hori,[1] Hideyuki Kobayashi,[1] Toshihiro Kawase,[2] Tetsuro Miyazaki,[2] Takahiro Kanno,[2] and Kenji Kawashima[2]

[1]Monozukuri Engineering Department, Tokyo Metropolitan College of Technology, 8-17-1 Minamisenjyu, Arakawa-ku, Tokyo 116-8523, Japan
[2]Institute of Biomaterials and Bioengineering, Tokyo Medical and Dental University, 2-3-10 Kanda-Surugadai, Chiyoda-ku, Tokyo 101-0062, Japan

(Received September 27, 2019; accepted October 28, 2019)

Keywords: visual interface, tele-operation, deep learning, laparoscope holder

We propose an interface for the tele-operation of a laparoscope-holder robot via head movement using facial feature point detection. Fourteen feature points on the operator's face are detected using a camera. The vertical and horizontal rotation angles and the distance between the face and the camera are estimated from the points using deep learning. The training data for deep learning are obtained using a dummy face. The root-mean-square error (RMSE) between the estimated and directly measured values is calculated for different numbers of nodes, layers, and epochs, and suitable numbers are determined from the RMSE values. The trained data are evaluated with four subjects. The effectiveness of the proposed method is demonstrated experimentally.

1. Introduction

In the tele-operation of a robot relying on indirect vision from a camera, the view angle of the camera affects the operation.[1–6] An interface to control the camera to provide a suitable view to the operator through head motion is effective. A laparoscope-holder robot named EMARO is used in minimally invasive surgery. The robot is operated by the head motion of the operator (a surgeon wearing a cap with a gyroscope). The pitch and yaw motions are controlled by the movement of the head while pushing the foot pedal.[7,8] However, the wiring of the gyroscope may interfere with the operation. The use of eye tracking for control is one solution. Interfaces that can be used to operate robots using eye tracking have been proposed.[9–14] However, controlling zoom in and out of the camera is difficult with eye tracking.

We have proposed an interface for the tele-operation of a camera, in which images of two markers attached to the operator's head are tracked by visual odometry.[15] However, misalignment of the markers leads to a risk of malfunction. To solve this problem, we previously proposed an interface for robotic tele-operation involving head movement and facial feature point detection. The feature points on the operator's face were detected using a camera

*Corresponding author: e-mail: minamoto@metro-cit.ac.jp
https://doi.org/10.18494/SAM.2020.2634

ISSN 0914-4935

and the position and posture of the face were calculated using these points.[16] Although the effectiveness of the interface has been experimentally demonstrated with a laparoscope holder, improvements in estimation accuracy are desired.

In this paper, we propose a method of improving the accuracy using deep learning. We set 14 feature points on the operator's face as the input for deep learning. The outputs are the horizontal and vertical rotation angles of the operator's head and the distance between the camera and the face. The training data for deep learning are obtained using a dummy face. The root-mean-square error (RMSE) between the estimated and directly measured values is calculated for different numbers of nodes, layers, and epochs, and optimal numbers are determined from the RMSE values. The rotations and the distance are estimated with four subjects.

2. Interface Using Detection of Facial Feature Points

2.1 Setup of interface

Figure 1 shows the setup of the interface. A USB camera (HD Pro Webcam C920, Logicool, USA) is installed on the monitor to capture the operator's face. The face image is transmitted to a PC (MB-W830S2-SSD2, Mouse Computer, JPN). Fourteen feature points on the face, as shown in Fig. 2, are detected using the software libraries of OpenCV and DLIB. The 14 feature points ($x_{1–14}$, $y_{1–14}$, $z_{1–14}$) are detected as the input for the deep neural network. Considering the application of the interface in surgery, the operator wears a mask. However, if we put some markers on the mask as feature points, our interface can be applied.

The horizontal and vertical rotations of the operator's head (θ_H and θ_V) and the distance between the camera and the face (L) are calculated using deep learning. The gazing point on the monitor is calculated from θ_H and θ_V. The point is displayed on the monitor as a red dot. The diameter of the dot is inversely proportional to L.

The monitor is divided into a 3 × 3 grid. The operator controls the laparoscope holder robot while gazing at the monitor. The robot remains in the same position when the dot is in the center area of the monitor. When the operator gazes at other squares of the grid, the robot

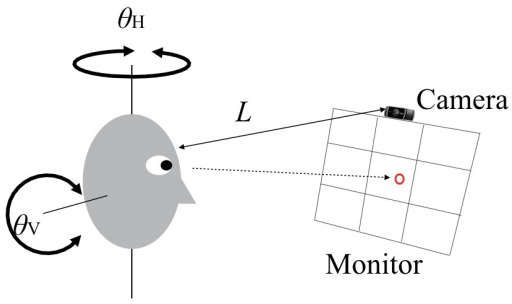

Fig. 1. (Color oneline) Setup of the interface.

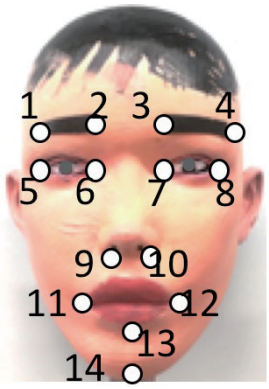

Fig. 2. (Color online) Feature points on operator's face.

moves at a constant velocity until the target object enters the center area. If the estimation accuracy of θ_H and θ_V is low, the intuitiveness of the operation deteriorates. Improving the accuracy of angle estimation will improve task efficiency. The initial distance between the monitor and the operator is 600 mm. The robot starts to zoom in when the distance becomes shorter than 550 mm and starts to zoom out when the distance becomes longer than 650 mm.

2.2 Estimation of face motion using deep learning

In this section, we explain the method of estimating the face motion from the 14 feature points using deep learning. We used a dummy face for the input of the deep learning. The height and width of the dummy face shown in Fig. 2 were 210 and 115 mm, respectively. The dummy face was sinusoidally rotated for horizontal and vertical directions at a frequency of 0.25 Hz. The rotational range was determined to be −30 to 30 deg for the horizontal direction and −25 to 25 deg for the vertical direction. The experiments were performed at L = 450, 500, and 550 mm. This is because the size of the dummy face is about 5/6 that of an adult face.

Figure 3 shows the structure of deep learning. Chainer (Preferred Networks, Inc.) was used for deep learning. The input data were 14 facial features detected during both horizontal and vertical motions, totaling 28 facial points. The output data were θ_H, θ_V, and L. The input and output data were recorded at 30 frames per second. Of the recorded data, 3600 are used as training data and 1200 are used as validation data after learning. Rectified linear unit (ReLU) and adaptive moment estimation (Adam) were used for the activation function and the optimizer. To learn the optimal weighting data by deep learning, it is necessary to set appropriate numbers of layers, nodes, and learning times. The optimal numbers of layers and nodes were determined from among 1, 2, 4, 8, and 16 layers and 10, 20, 30, and 50 nodes

Figure 4 shows the calculated θ_H and θ_V for different numbers of nodes with two layers. The horizontal axis shows the number of nodes and the vertical axis shows the root of square error (RSE) for the calculated angles. The calculation error is minimized with 30 nodes. The same tendency can be observed for different numbers of layers. Figure 5 shows the calculated results of L for the different numbers of nodes with two layers. The horizontal axis shows the number of nodes and the vertical axis shows the RSE for the calculated length. We determined the optimal number of nodes to be 30 from the calculated results.

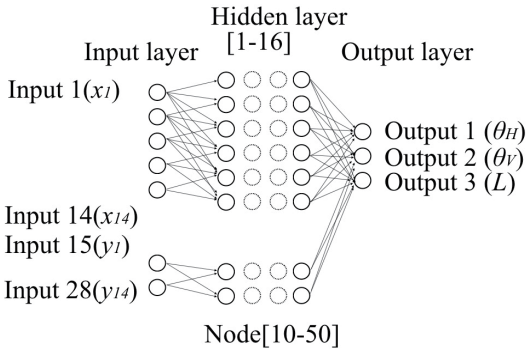

Fig. 3. Structure of deep learning.

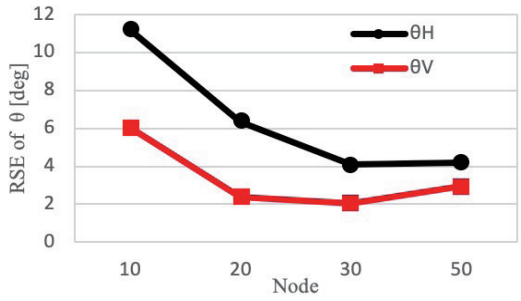

Fig. 4. (Color online) RSE values of angles for different numbers of nodes.

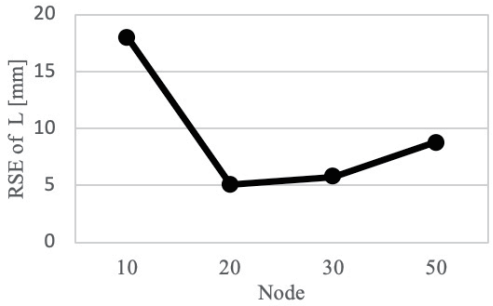

Fig. 5. RSE values of length for different numbers of nodes.

Figures 6 and 7 respectively show the RSEs of the angles and the length calculated for 1, 2, 4, 8, and 16 layers with 30 nodes. The horizontal axis shows the number of layers and the vertical axis shows the RSE for the calculated values. We determined the optimal number of layers to be two from the calculated results. Figure 8 shows the learning situation. Since the accuracy peaked when the number of epochs approached 500, this number of epochs was selected. The training in deep learning was performed with 3600 data values using 30 nodes, 2 layers, and 500 epochs. Then, we performed the validation using 1200 data values.

Figure 9 shows the estimated results of θ_H for L = 300 mm. The same distance as in Ref. 16 was selected for comparison. The RMSE was 4.36 deg. In Ref. 16, we obtained the position and posture of the face by solving the perspective-*n*-point (PnP) problem. The results are shown in Fig. 10. The RMSE was 6.42 deg. The phase delay observed in Fig. 10 is due to the computation time to solve the PnP problem. In Fig. 9, the phase delay cannot be observed because the computation time is less than 1 ms. It is clear that the proposed method is more accurate by comparing Figs. 9 and 10.

3. Experiments with Interface

Four subjects were tested with the interface trained using the dummy face, as described in Sect. 2.

3.1 Experimental procedure

The proposed interface shown in Fig. 1 was tested with four subjects using the data trained with deep learning. A gyro sensor (MPU9250/6500, HiLetgo, China) was mounted on the subject's head as a reference for the estimated angles. The experiments were performed with the lengths of 550, 600, and 650 mm.

The subjects were asked to rotate their face horizontally by −10, −20, 10, and 20 deg while watching the output of the gyro sensor from the initial posture facing forward. The clockwise motion was set to be the positive direction. Then, the subjects rotated their face vertically by −10, −20, 10, and 20 deg, where the upward direction was set to be the positive direction.

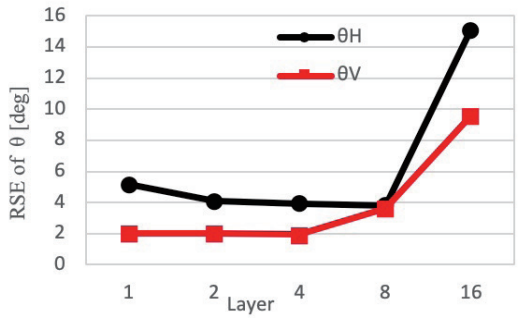

Fig. 6. (Color online) RSE values of angles for different numbers of layers.

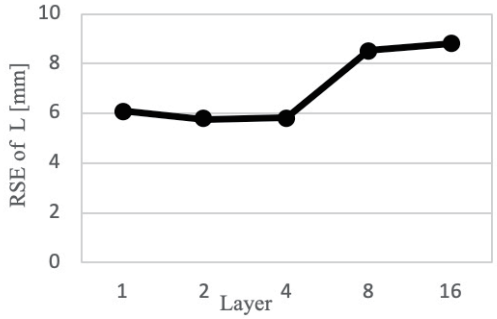

Fig. 7. RSE values of length for different numbers of layers.

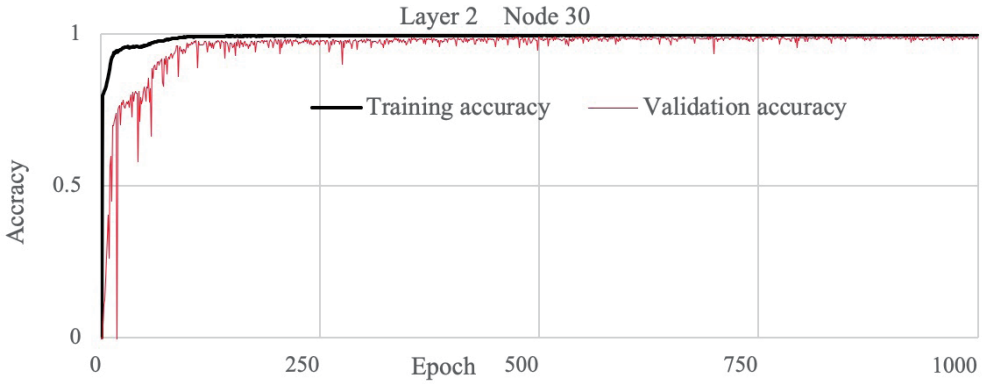

Fig. 8. (Color online) Learning situation.

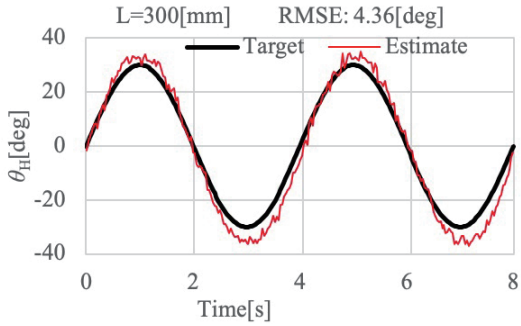

Fig. 9. (Color online) RMSE value of θ_H obtained using deep learning.

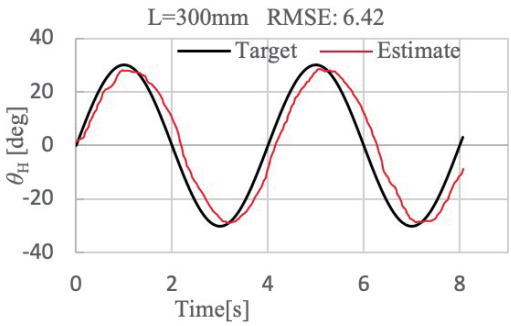

Fig. 10. (Color online) RMSE value of θ_H obtained by solving PnP problem.

3.2 Experimental results

Figure 11 shows the experimental results of the horizontal rotation for lengths of 550, 600, and 650 mm. The average RMSE of four subjects was plotted on the longitudinal axis. The value was minimum when L = 600 mm. Figure 12 shows the experimental results of the vertical rotation for the three different lengths. The estimation accuracy was high for the

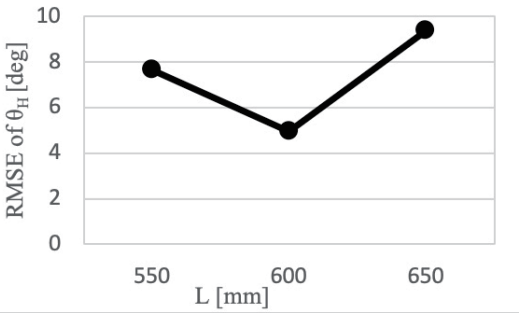

Fig. 11. RMSE values of horizontal rotation for different lengths.

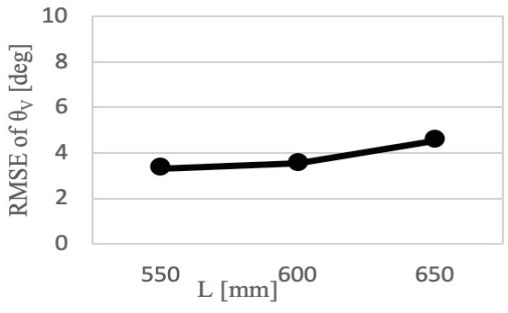

Fig. 12. RMSE values of vertical rotation for different lengths.

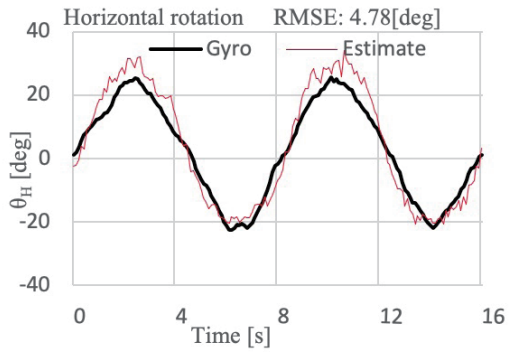

Fig. 13. (Color online) Experimental results of horizontal rotation.

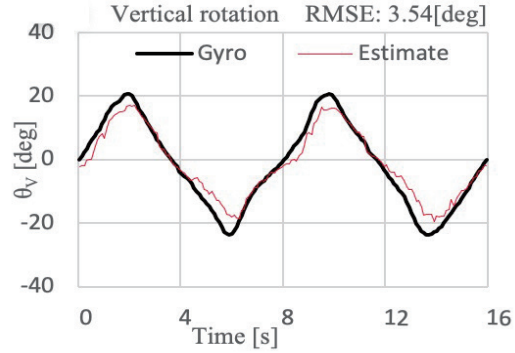

Fig. 14. (Color online) Experimental results of vertical rotation.

vertical rotation. This is considered to be due to the arrangement of the feature points. The heights of the feature points 2–14 were larger than the widths of points 1–4. Since the height was larger than the width, the feature points were markedly changed upon vertical rotation of the face. Therefore, the estimation accuracy was higher in the vertical rotation than in the horizontal rotation.

Next, the subjects performed sinusoidal rotational motion. The initial length from the camera was 600 mm. The subjects rotated their face by about 20 deg while watching the output of the gyro sensor. Figures 13 and 14 respectively show the experimental results of the horizontal and vertical rotations. The black line shows the output of the gyro sensor. The red line shows the results estimated from the feature points of the face using deep learning. The RSE values were 4.78 and 3.53 deg, respectively. The calculation delay resulting from the time taken to solve the problem[16] was about 75 ms compared with only 25 ms for the proposed method with deep learning. There was a slight phase delay in the vertical rotation. The dummy face was smaller than the subject's face; therefore, the detection accuracy of the feature points on the chin was low. The accuracy can be improved by changing the size of the dummy face. However, the black and red lines are in good agreement. The effectiveness of the proposed method was confirmed from the experimental results.

4. Conclusions

In this paper, we proposed an interface for the tele-operation of robots via head movement using facial feature point detection. The vertical and horizontal rotation angles and the distance between the face and the camera were estimated from 14 feature points on the operator's face using deep learning. The training data for deep learning were obtained using a dummy face. The RMSE between the estimated values and the values directly measured using sensors was calculated for different numbers of nodes, layers, and epochs. We found that 2 layers, 30 nodes, and 500 epochs were suitable for deep learning. The trained data were evaluated with four subjects. We confirmed that sinusoidal head motion is effective for training in the proposed method.

Acknowledgments

This research is based on the Cooperative Research Project of Research Center for Biomedical Engineering.

References

1 J. Y. Chen, E. C. Hass, and M. J. Barnes: IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **37** (2007) 1231. https://doi.org/10.1109/TSMCC.2007.905819
2 K. Hoshino, M. Kitani, R. Asami, N. Sato, Y. Morita, T. Fujiwara, T. Endo, and F. Matsuno: 2018 IEEE Int. Conf. Intelligence and Safety for Robotics (2018) 18236248. https://doi.org/10.1109/IISR.2018.8535854
3 K. Chayama, A. Fujioka, K. Kawashima, H. Yamamoto, Y. Nitta, C. Ueki, A. Yamashita, and H. Asama: J. Rob. Mechatron. **26** (2014) 403. https://doi.org/10.20965/jrm.2014.p0403
4 J. Yang, M. Kamezaki, R. Sato, H. Iwata, and S. Sugano: 2015 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IEEE, 2015) 15667007. https://doi.org/10.1109/IROS.2015.7354135
5 N. Marturi, A. Rastegarpanah, C. Takahashi, M. Adjigble, R. Stolkin, S. Zurek, M. Kopicki, M. Talha, J. A. Kuo, and Y. Bekiroglu: 2016 Int. Conf. Robotics and Automation for Humanitarian Applications (IEEE, 2016) 16902272. https://doi.org/10.1109/RAHA.2016.7931866
6 T. Sasaki and K. Kawashima: Autom. Constr. **17** (2008) 226. https://doi.org/10.1016/j.autcon.2008.02.004
7 Y. Matsuoka, K. Kihara, K. Kawashima, and Y. Fujii: Videosurgery and Other Miniinvasive Techniques **9** (2014) 613. https://doi.org/10.5114/wiitm.2014.44135
8 K. Tadano and K. Kawashima: Int. J. Med. Rob. Comput. Assisted Surg. **11** (2015) 331. https://doi.org/10.1002/rcs.1606
9 P. D. S. H. Gunawardane, N. T. Medagedara, and B. G. D. A. Madhusanka: 2015 8th Int. Conf. Ubi-Media Computing (IEEE, 2014) 15523379. https://doi.org/10.1109/UMEDIA.2015.7297437
10 F. Fanfani, G. Monterossi, A. Fagotti, C. Rossitto, S. Gueli Alletti, B. Costantini, V. Gallotta, L. Selvaggi, S. Restaino, and G. Scambia: Surgi. Endos. **30** (2016) 215. https://doi.org/10.1007/s00464-015-4187-9
11 D. Zhu, T. Gedeon, and K. Taylor: Interact. Comput. **23** (2011) 85. https://doi.org/10.1016/j.intcom.2010.10.003
12 Y. Suzuki, K. Shirahada, M. Kosaka, and A. Maki: Int. Conf. Services Systems and Services Management (ICSSSM) (2012) 12905743. https://doi.org/10.1109/ICSSSM.2012.6252344
13 Y. Cao, S. Miura, Y. Kobayashi, K. Kawamura, S. Sugano, and M. G. Fujie: IEEE Rob. and Autom. Lett. **1** (IEEE, 2016) 531. https://doi.org/10.1109/LRA.2016.2521894
14 M. Minamoto, Y. Suzuki, T. Kanno, and K. Kawashima: 2017 Int. Conf. Mechatronics and Automation (IEEE, 2017) 17137216. https://doi.org/10.1109/ICMA.2017.8016122
15 M. Minamoto, M. Sato, T. Kanno, and K. Kawashima: 2018 IEEE Int. Conf. Mechatronics and Automation (IEEE, 2018) 18149355. https://doi.org/10.1109/ICMA.2018.8484620
16 M. Minamoto, H. Sato, T. Kanno, T. Miyazaki, T. Kawase, and K. Kawashima: 2019 IEEE Int. Conf. Mechatronics and Automation (IEEE, 2019) 18956807. https://doi.org/10.1109/ICMA.2019.8816385

About the Authors

Masahiko Minamoto received his doctoral degree in engineering from the Faculty of Information Science and Electrical Engineering at Kyushu University in 2001. From 1990 to 2001, he worked as a researcher at Fujita Corporation. From 2004 to 2006, he worked as a professor at Kindai University Technical College. Since 2006, he has been a professor at Tokyo Metropolitan College of Technology. His research interests are in medical robotics and operational interfaces. (minamoto@metro-cit.ac.jp)

Shigeki Hori received his doctoral degree in engineering from the Department of Mechanical Sciences and Engineering at Kanagawa Institute of Technology in 2000. From 2000 to 2002, he worked as a researcher at in the Satellite Venture Business Laboratory, University of Electro-Communications. Since 2004, he has been an associate professor at Tokyo Metropolitan College of Technology. His research interests include force control, multiobjective design, robotics, and applications.
(hori@metro-cit.ac.jp)

Hideyuki Kobayashi is studying robotics at Tokyo Metropolitan College of Technology.

Toshihiro Kawase received his B.S., M.S., and Ph.D. degrees from Tokyo Institute of Technology, Tokyo, Japan, in 2007, 2009, and 2012, respectively. He was a research fellow at Research Institute of National Rehabilitation Center for Persons with Disabilities from 2012 to 2015, and worked as a postdoctoral fellow and a specially appointed assistant professor at Tokyo Institute of Technology from 2015 to 2017. He is currently an assistant professor at Tokyo Medical and Dental University and Tokyo Institute of Technology. His research interests include rehabilitation robotics, medical robots, and biological signal processing. (kawase.bmc@tmd.ac.jp)

Tetsuro Miyazaki received his doctoral degree in engineering from the Department of Mechanical Sciences and Engineering at Tokyo Institute of Technology in 2014. From 2014 to 2017, he worked as a research assistant (2014 to 2015) and an assistant professor (2015 to 2017) at Yokohama National University. Since April 2017, he has been an assistant professor at the Institute of Biomaterials and Bioengineering at Tokyo Medical and Dental University. His research interests are in mechanical engineering, control engineering, power assistive devices, and medical welfare robotics.
(tmiyazaki.bmc@tmd.ac.jp)

Takahiro Kanno received his doctoral degree in engineering from the Department of Mechanical Engineering and Science at Kyoto University in 2013. In 2013, he worked as a postdoctoral researcher in the Precision and Intelligence Laboratory, Tokyo Institute of Technology. From 2013 to 2019, he worked as an assistant professor at Tokyo Medical and Dental University. Since June 2019, he has been an associate professor there. His research interests are in medical robotics, control engineering, and tele-operation. (kanno.bmc@tmd.ac.jp)

Kenji Kawashima received his doctoral degree in engineering from the Department of Control Engineering at Tokyo Institute of Technology in 1997. From 1997 to 2000, he worked as a research assistant at Tokyo Metropolitan College of Technology. He then worked as an associate professor in the Precision and Intelligence Laboratory at Tokyo Institute of Technology. Since April 2013, he has been a professor at the Institute of Biomaterials and Bioengineering at Tokyo Medical and Dental University. His research interests are in medical robotics, control engineering, fluid measurement, and control. (kkawa.bmc@tmd.ac.jp)