Proceedings of 2021 IEEE
International Conference on Mechatronics and Automation
August 8 - 11, Takamatsu, Japan

Evaluation of a Reinforcement Learning Algorithm for Vascular Intervention Surgery

Fanxu Meng [1], Shuxiang Guo [1,2*], Wei Zhou [1], Zhengyang Chen [1]

[1] *Key Laboratory of Convergence Biomedical Engineering System and Healthcare Technology, The Ministry of Industry and Information Technology, School of Life Science, Beijing Institute of Technology, No.5, Zhongguancun South Street, Haidian District, Beijing 100081, China*

[2] *Faculty of Engineering, Kagawa University, 2217-20 Hayashi-cho, Takamatsu, Kagawa 760-8521, Japan*

E-mails: {mengfanxu & guoshuxiang & zhouwei & chenzhengyang}@bit.edu.cn;

* Corresponding author

***Abstract* – With the increasing use of vascular interventions, catheter navigation in complex vessels has become even more critical. Vascular intervention surgeries also require more precise manipulation and a more intelligent system to ensure the safety of the patients. In this paper, a virtual training model based on deep reinforcement learning was designed to navigate the catheter into the aortic arch. The whole experiment was carried out in a virtual environment, and a reinforcement learning method was used to test the performance of catheter autonomous navigation in vessels. Finally, the model was successfully trained and results were analyzed basing on previous work. The results obtained would be more convincing if the model was more complex and closer to the actual vessels.**

Index Terms -Vascular interventional surgery, Reinforcement learning, Catheter navigation.

I. Introduction

With the rapid increase in cardiovascular morbidity, minimally invasive vascular interventions have rapidly replaced traditional open or cranial surgery due to their patient-friendly nature [1]. For many years, medical robots have been used in surgery and healthcare, and the use of robots in surgery has been beneficial for departments such as head and neck, cardiac and urology Robotics in cardiac surgery [2].

The ultimate goal of vascular intervention is to perform the procedure without compromising the integrity of the chest. Catheter access to the heart can be less traumatic for the patient [3]. However, it significantly increases the complexity of the surgical approach, requiring more sophisticated instruments, greater precision, dexterity and intuitive remote manipulation. It also has its disadvantages: interventional procedures are highly dependent on the surgeon's surgical experience, and the cost of training qualified surgeons is high [4].

Existing master-slave interventional robots are passive recipients of the surgeon's actions. In the more mature CorPath GRX system, autonomy is also limited to compensation for the angle of rotation of the guidewire and does not tap into multidimensional information [5]. There is, therefore, a need to combine simulation technology and in vitro physical vascular models to build a more intelligent robotic system for vascular interventions.

In recent years, automation technologies based on deep learning and reinforcement learning have been rapidly shaped and implemented, showing a high scientific value. In the field of surgical robots, the automated and intelligent operation of surgical instruments based on ensuring surgical safety has been a goal pursued by researchers. However, compared to general artificial intelligence systems, surgical data is costly to acquire and less interpretable, and a model is only valid for a single surgical task and lacks generalization capability [6]. However, in contrast to the complex multi-process decision-making tasks such as resection and suturing in general surgery, the simple, intravascular dynamic and stable mode of operation of tube-filament access in interventional surgery provides a good environment for the training of automatic models and makes it easier to exploit the advantages of high precision and stability of robotic control [7]. The study of critical technologies for the automatic control of surgical robotic tube and wire access is a realistic and scientific attempt. The simplification and automation of tube and wire access operations can greatly relieve the technical and experiential pressure on interventional surgeons and provide new perspectives and tools for the development of surgical robots and vascular surgery by quantifying the implicit intraoperative features.

Researchers from different teams have experimented with virtual training systems, automated catheter navigation in vascular in vitro models. Yang et al. at Imperial College of Technology attempted to use reinforcement learning algorithms to control a vascular interventional robot to complete autonomous over-arch operations on four aortic arch models, showing that the robot's operating force fluctuated over a significantly lower range than that of a human hand and operated at approximately half the speed of a human hand [8]. Using the dueling deep Q-learning (DQN) algorithm to control catheter entry into a heart model, You et al. at the University of Ulsan also demonstrated that a reinforcement learning strategy based on a simulated environment could control an actual catheter to complete a cardiac entry [9]. However, there are still problems such as lack of accuracy and a relatively simple model. Initial control using the Deep Deterministic Policy Gradients (DDPG) algorithm was implemented in a 2D environment by Karstensen et al. at Fraunhofer IPA, Germany, and performed well in a planar vascular model, but fell short for higher-level path control [10].

In this paper, we use reinforcement learning algorithms to implement control of catheter access in a simulation engine.

978-1-6654-4098-1/21/$31.00 ©2021 IEEE

The simulation process performed is an over-arching operation of the aortic arch, which is eventually trained successfully in a virtual environment and can be used in subsequent catheter access navigation in a real environment.

The paper is structured as follows: Section II focuses on the building of vascular models and reinforcement learning methods. Section III introduces the arrangements of the experiment and analyse the results. Section IV and V present the discussion and the summarization of the paper.

II. Methods

The methods we use in this research includes four main parts: modelling, reinforcement learning, the specific algorithm we use and the training engine. The mentioned four parts cover the two main issues- algorithms and simulation environments for reinforcement learning.

A. Modeling

The process of modelling the aortic arch and catheter progresses from the base shape to the inclusion of more features. Its medical characteristics are primarily realistic, and the established model is subsequently fed into the Unity engine to establish its environmental parameters.

The movement of the guidewire through the vascular tree is simulated using Unity [11]. The walls of the vascular tree are rigid. The lumen is empty; thus, no dynamic resistance to the guidewire motion is considered.

Our model, as shown in Fig. 1, is based on the angiographic image of the aortic arch, but with a partial simplification of the vascular connection at the vessel cross-section, so that the cross-sections of the model vessels are all circular, while the interface is a smooth connection.

In order to verify the function of reinforcement learning, this paper obtains the accuracy and stability of catheter autonomy learning by modeling simulated over-arch operations on the aortic arch.

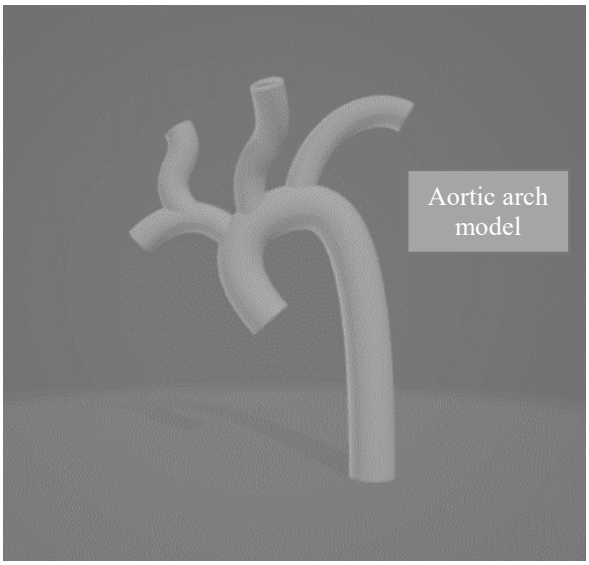


Fig. 1 Preliminary modelling of the aortic arch.

B. Reinforcement Learning Methods

The basic model of reinforcement learning is the individual-environment interaction. The individual/intelligent agent is the part of the individual that can take a series of actions and expects to achieve a high benefit or goal. The other parts associated with this are referred to as the environment. The whole process is discretized into different time steps. At each moment, the environment and the individual interact accordingly. The individual can take specific actions, which are imposed on the environment. After receiving the individual's action, the environment gives the individual feedback on the current state of the environment and on the reward that has been generated as a result of the previous action [12].

Reinforcement learning is a formal framework (Fig. 2) that uses Markov decision processes to define the process by which a learning intelligence interacts with its environment using states, actions and gains.

In the basic setup of reinforcement learning, there are essential elements such as agent, environment, action, state, reward, etc. The agent interacts with the environment to generate trajectories, and by performing the action, the environment changes its state. The agent interacts with the environment, generating trajectories that cause the environment to change state by performing an action; the environment then gives the agent a reward (positive or negative) for its current action. Through this interaction, more and more experience is accumulated, and the policy is updated to finally form a closed loop. The mystery of why reinforcement learning can model the long-term benefits of decision-making lies in its optimization goal. To be specific, at each moment, the reward is a specific value and the agent's goal is to maximize the expectation of the reward it obtains [13]. This means that instead of maximizing the immediate reward, it maximizes the cumulative reward over time.

For each state in a discrete finite state space, the probability of its transfer to another state depends only on that state itself- what is referred to as Markovness- the Markov decision process (MDP) differs from a Markov chain in that in each state, an action can be chosen by the individual from the space of possible actions [14]. Also, Markovness can generally be enhanced if not just the current moment's state, but multiple previous moments' states are selected and superimposed as the current state.

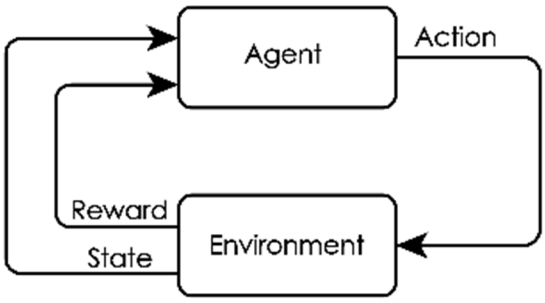


Fig. 2 Reinforcement learning method process.

C. *Asynchronous Advantage Actor-critic(A3C)*

In this section, the background of the emergence of A3C and its advantages are presented with a contrast between previous methods. Finally, the single A3C network algorithm is given.

Asynchronous advantage actor-critic is an algorithm proposed by Google DeepMind to solve the Actor-Critic non-convergence problem [15]. While DQN is vital because it has an experienced pool that reduces the correlation between data, A3C proposes an alternative way to reduce the correlation between data: asynchronously.

A3C creates multiple parallel environments and allows multiple agents with sub-structures to update parameters in the main structure on these parallel environments simultaneously. The agents in parallel do not interfere with each other, while the parameter updates of the primary structure are interfered with by the discontinuity of the updates submitted by the substructures, so the correlation of the updates is reduced, and convergence is improved.

The main idea of A3C is asynchronous, corresponding to the asynchronous distributed RL framework. Corresponding to Google's Gorilla platform Massively Parallel Methods for Deep Reinforcement Learning in 2015, Gorilla uses different machines with the same PS. While in A3C, it is the same machine with multi-core CPUs, which reduces the parameter, and in A3C, it is the same machine with multiple CPUs, which reduces the cost of transferring parameters and gradients, and the validation iterations are significantly faster in the paper. And more importantly, it is an actor-learner pair with multiple threads on the same machine; each thread corresponds to a different exploration policy, and the overall inter-sample correlation is low, so it is no longer necessary to introduce an experience replay mechanism in DQN for training [16]. This enables an on-policy approach to training. In addition, the CPU is used in training instead of the GPU because the RL batch is generally small during training and the GPU is much idle while waiting for new data.

Different types of deep neural networks provide an efficient operational representation of the policy optimization task in DRL [17]. To alleviate the instability that occurs when combining traditional policy gradient methods with neural networks, various types of deep policy gradient methods use an empirical replay mechanism to eliminate the correlation between training data.

However, there are two main problems with the empirical replay mechanism: Each real-time interaction between the agent and the environment requires a lot of memory and computational power; the experience replay mechanism requires the agent to learn using an off-policy approach, which can only be updated based on the data generated by the old policy; and the training of DRLs has previously relied on computationally powerful graphics processors [18].

The A3C algorithm first constructs a global network. This network will consist of convolutional layers for spatial dependencies, followed by LSTM layers for temporal dependencies, and finally value and policy output layers [19]. The process of the algorithm is shown below:

Algorithm 1. A3C network learning process

1 **Input**: public part of the A3C neural network parameters θ, ω
2 Update time series t=1
3 Reset the gradient updates of Actor and Critic
4 $\theta'=\theta$, $\omega'=\omega$
5 Initialize state start t
6 Choose action a_t based on policy $\pi(a_t| s_t; \theta)$
7 Execute action a_t to get reward r_t and new state
8 $t\leftarrow t+1, T\leftarrow T+1$
9 If s_t is terminated, then go to step 10, otherwise go back to step 6
10 Calculate Q(s,t) for the last time series position s_t
11 For $i\in(t-1,t-2,\dots .t_{start})$:
1) Calculate Q(s, i) for each moment: $Q(s,i)=r_i+\gamma Q(s,i+1)$
2) Local gradient update of the cumulative Actor
3) Local gradient update of the cumulative Critic
12 Update the model parameters of the global neural network.
$\theta=\theta-\alpha d\theta$, $w=w-\beta d\omega$
13 If $T>T_{max}$, then the algorithm ends and **outputs** the public part of the A3C neural network parameters θ, ω, otherwise go to step 3

D. *Training Engine*

The setup consists of the controller (DRL-Agent), a simulated environment created in Unity. The simulation in Unity is used to generate training data for the DRL-Agent [20].

In addition, Unity's ml-agents provide a reinforcement learning environment in which the vascular rigidity model can also be modelled in the unity environment and trained for entry to verify the applicability of the algorithm. Unity's setup is simple among the simulation applications, but the adaptation to vascular elasticity is somewhat lacking when used [21].

Most of the current games have a large number of Unity games, a perfect engine, and a good training environment to build. Since Unity can be cross-platform, it can be trained under Windows and Linux platforms and then converted to WebGL for publishing to the web [22]. Furthermore, ml-agents is an open-source plug-in for Unity, which allows developers to train in Unity's environment, without even writing code in python, without a deep understanding of PPO, SAC and other algorithms [23]. As long as developers configure the parameters, they can easily use reinforcement learning algorithms to train their own models [24].

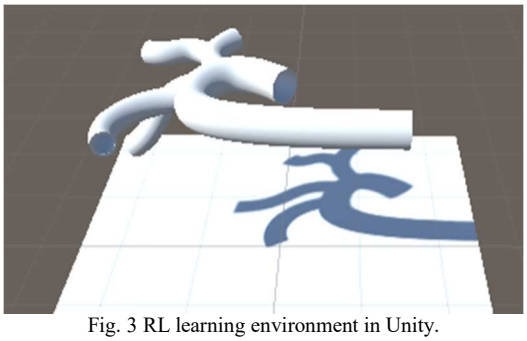


Fig. 3 RL learning environment in Unity.

III. Experiments and Results

TABLE I
Parameters of the Training Process

Simulation distribution update parameters	
Minibatch size	2000
Training update size	1000
Replay memory size	400000
Update frequency	20
Learning rate	0.001
Discount factor	0.9
Episode size	2000
Episode step	1000
Optimizer	RMSPRob

The parameters of the reinforcement learning process were set as Table 1. And the results of A3C learning are shown in Fig. 5, and as shown in the figure, the final results obtained from the training can corroborate that the model was successfully trained to achieve stable returns, with large fluctuations in the choice of losses, which may be related to the instability of the model itself at the time of collision.

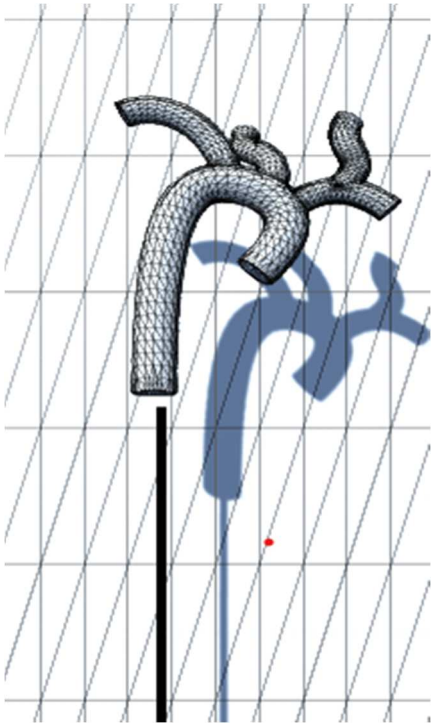
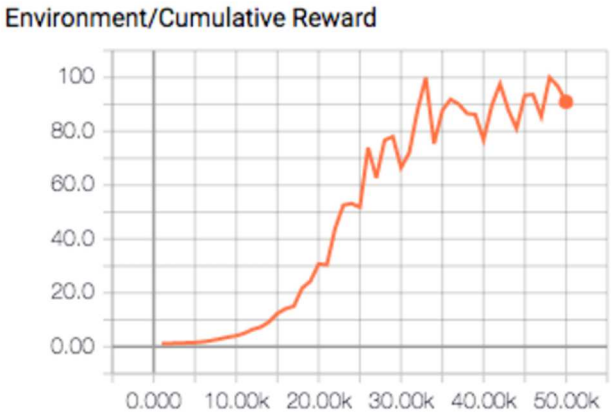
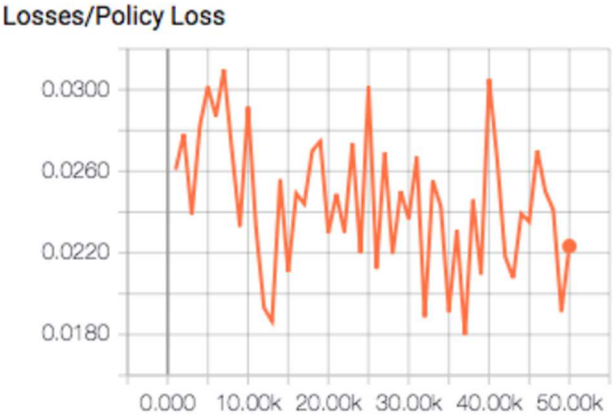


Fig. 4 RL learning process of catheter in Unity.

The A3C algorithm was added and retrained for comparison to obtain different results of the algorithm for the simulation of vascular interventional procedures in the built training environment. The virtual environment for training is represented in Fig. 4, and using the catheter to access the aortic arch and perform over-arch manipulation, we derived the subsequent results.



(a) Result of the cumulative reward.



(b) Result of the policy loss.

Fig.5 Results using A3C algorithms.

From Fig. 5(a), it can be concluded that the training of reinforcement learning starts to stabilize when it reaches 30K times, and the final reward value reaches stability in 30K-50K training episodes, but there are still fluctuations. But the overall result of the reward is consistent with the expected training results of the A3C algorithm.

From Fig. 5(b), it can be seen that the loss function does not get a convergence result after a certain time of training, and always fluctuates up and down throughout the training process, the magnitude of fluctuations is larger, but the overall trend is converging to zero. This is in line with the characteristics and trend of policy loss in reinforcement learning training.

IV. Discussions

The research in this paper is divided into two main processes: the modelling part and the model training part.

In the modelling part, we try to provide an environment for subsequent algorithm training by modelling the aortic arch. The model of the aortic arch is based on the real vascular environment, but due to the high complexity of the vasculature and the difficulty to monitor the internal environment in real time, the modelling process simplifies the structure of the vasculature and ignores the blood flow and the more complex respiration and pulsation in the vasculature, which needs to be

solved by more accurate modelling and imposing fluid motion in the model.

In the training part of the model, the final training results obtained using the A3C algorithm showed that the reward training of the model was more in line with expectations and was able to achieve reward stability in a short step; however, the fluctuation of the loss function was large and did not converge after a certain length of training, which may be due to the parameter settings in the training, or the large fluctuation of the catheter position in the training environment This may be due to the parameter settings in the training, or it may be due to the large fluctuations in the position of the conduit in the training environment, which does not converge to the same stable path. These problems need to be solved by tuning the parameters and imposing more constraints on the catheter in subsequent studies.

V. CONCLUSIONS

In this paper, we aim to use reinforcement learning algorithms to implement control of catheter access in a simulation engine. The simulation process performed is an over-arching operation of the aortic arch, which is eventually trained successfully in a virtual environment and can be used in subsequent catheter access navigation in a real environment. As the result shows, this concept is feasible and can further improve the model accuracy and algorithm accuracy.

To date, we have been investigating the feasibility of our approach using highly idealized vascular models and just a guidewire. But there is still a long way to go before reinforcement learning can be applied to real-world scenarios. The complex environment that changes at any time in the vasculature requires a simulation environment with sufficient vascular complexity, while being able to simulate blood flow, pulse, heartbeat and other influencing factors, which will be a huge project. The results show that the A3C algorithm is able to obtain more desirable results in these idealized models. Future studies should address the mentioned limitations by adjusting the agents and settings to fit more realistic vessel geometries, as well as by using both guidewires and catheters. Also, the specific contact of the vessel wall with the catheter guidewire and its own elastic characteristics should also be properly characterized.

ACKNOWLEDGEMENT

This work was supported in part by National High-tech Research and Development Program (863 Program) of China (2015AA043202).

REFERENCES

[1] S. Guo, et al, “Machine learning-based Operation Skills Assessment with Vascular Difficulty Index for Vascular Intervention Surgery,” *Medical & Biological Engineering & Computing*, vol. 58, no.8, pp. 1707- 1721, 2020.

[2] Swaminathan R V, Rao S V, “Robotic-Assisted Transracial Diagnostic Coronary Angiography,” *Catheterization & Cardiovascular Interventions*, vol. 10, no. 1, pp. 7-21, 2018.

[3] Y. Zhao, et al, “A novel noncontact detection method of surgeon’s operation for a master-slave endovascular surgery robot,” *Medical & Biological Engineering & Computing*, vol. 58, no. 4, pp. 871-885,2020.

[4] Y. Wang, et al, “Online Measuring and Evaluation of Guidewire Inserting Resistance for Robotic Interventional Surgery Systems,” *Microsystem Technologies*, vol. 24, no. 8, pp. 3467-3477, 2018.

[5] Britz. G, et al, “Neuro-endovascular-specific engineering modifications to the CorPath GRX robotic system,” *Journal of neurosurgery*, vol. 1, no. 1, pp. 1-7, 2019.

[6] C. Yang, et al, “A Vascular Interventional Surgical Robot Based on Surgeon's Operating Skills,” *Medical & Biological Engineering & Computing*, vol. 57, no. 9, pp. 1999-2010, 2019.

[7] J. Guo, et al, “A Vascular Interventional Surgical Robotic System based on Force-Visual Feedback,” *IEEE Sensors Journal*, vol.19, no. 23, pp. 11081-11089, 2019.

[8] Chi. W, et al, “Trajectory optimization of robot-assisted endovascular catheterization with reinforcement learning,” *IEEE International Conference on Intelligent Robots and Systems,* pp. 3875-3881,2018.

[9] Hyeonseok, et al. “Automatic control of cardiac ablation catheter with deep reinforcement learning method,” *Journal of Mechanical Science and Technology*, vol. 33, no. 11, pp. 5415-5423, 2019.

[10] Karstensen. L, et al. “Autonomous guidewire navigation in a two-dimensional vascular phantom,” *Current Directions in Biomedical Engineering*, vol. 6, no. 1, 2020.

[11] K. Wang, et al, “Endovascular intervention robot with multi-manipulators for surgical procedures: Dexterity, adaptability, and practicability,” *Robotics and Computer-Integrated Manufacturing*, vol. 56, pp. 75-84, 2019.

[12] J. Kim, et al, “A novel tip-positioning control of a magnetically steerable guidewire in sharply curved blood vessel for percutaneous coronary intervention,” *International Journal of Control, Automation and Systems*, vol. 17, no. 8, pp. 2069-2082, 2019.

[13] Y. Zhao, et al, “A CNNs-based Prototype Method of Unstructured Surgical State Perception and Navigation for an Endovascular Surgery Robot,” *Medical & Biological Engineering & Computing*, vol. 57, no. 9, pp. 1875-1887, 2019.

[14] Khan E M, et al, “First experience with a novel robotic remote catheter system: Amigo™ mapping trial,” *The Journal of Interventional Cardiac Electrophysiology*, vol.37, no.2, pp.121-129, 2013.

[15] X. Yin, et al, “Safety operation consciousness realization of MR fluids-base novel haptic interface for teleoperated catheter minimally invasive neuro surgery,” *IEEE/ASME Transactions on Mechatronics,* vol. 21, no. 2, pp. 1043-1054, 2015.

[16] Y. Wang, et al, “Design and evaluation of safety operation VR training system for robotic catheter surgery,” *Medical & Biological Engineering & Computing*, vol.56, no.1, pp.25-35, DOI 10.1007/s11517-017-1666-2, 2017.

[17] S. Guo, et al, “A Novel Robot-Assisted Endovascular Catheterization System with Haptic Force Feedback,” *The IEEE Transactions on Robotics*, vol. 35, no. 3, pp. 685-696, 2019.

[18] Y. Wang, et al, “Surgeons Operation Skill-Based Control Strategy and Preliminary Evaluation for a Vascular Interventional Surgical Robot,” *Journal of Medical and Biological Engineering*, vol. 39, no. 5, pp. 653-664, 2019.

[19] Y. Zhao, et al, “A CNNs-based Prototype Method of Unstructured Surgical State Perception and Navigation for an Endovascular Surgery Robot”, *Medical & Biological Engineering & Computing*, vol.57, no.9, pp.1875-1887, DOI:10.1007/s11517-019-02002-0, 2019.

[20] X. Bao, et al, “Compensatory force measurement and multimodal force feedback for remote-controlled vascular interventional robot,” *Biomedical Microdevices*, vol.20, no.3, pp.74.1-74.11, DOI: 10.1007/s10544-018-0318-0, 2018.

[21] Y. Zhao, et al, “Operating force information on-line acquisition of a novel slave manipulator for vascular interventional surgery,” *Biomed Microdevices*, vol. 20, no. 2, pp. 1-13, 2018.

[22] L. Zhang, et al. “Design and performance evaluation of collision protection-based safety operation for a haptic robot-assisted catheter operating system,” *The Journal of Biomedical Microdevices*, vol.20, no.2, pp. 22-36, 2018.

[23] X. Jin, et al, “Development of a Tactile Sensing Robot-assisted System for Vascular Interventional Surgery”, *IEEE Sensors Journal*, vol.21, No.10, pp.12284-12294, DOI:10.1109/JSEN.2021.3066424, 2021.

[24] Y. Song, et al, “Performance evaluation of a robot-assisted catheter operating system with haptic feedback”, *Biomedical Microdevices*, Vol.20, No.2, DOI: 10.1007/s10544-018-0294-4, 2018.