Proceedings of the 5th International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/2027.39075

Visual-Tactile Cooperative Control for Dynamic Soft Tissue Localization

Jingzhan Hu

Sino-British College, University of Shanghai for Science and Technology, Shanghai, China
sbc-23-8155@sbc.usst.edu.cn

Abstract. In robot minimally invasive surgery, soft tissue keeps moving. It moves because of breathing, heartbeat, and when tools pull on it. If the control system is just designed to correct only one target position it cannot track the tissue for an extended period of time. This paper looks at three kinds of studies. The first is visual localization without markers, the second is tactile sensing at the end of the instrument, and the third is how vision and touch work together. The paper also compares three types of control methods, which are vision-led, touch-led, and fusion at information level or control level. The reviewed studies indicate that vision is good at telling where the target is and where the instrument is. Touch is good at telling the force after the instrument touches tissue, and when contact happens. The two are not the same thing, and they cannot just replace each other. When they work together, people still need to think about time delay, who is in control, and what to do when a sensor fails. Right now, not many studies test these things in a real clinical environment, or test what happens when one sensor breaks. Later, people should use tissue phantoms, animal tests, and real patients to check how big the position error is, how big the contact force is, and what the system does when something goes wrong.

Keywords: minimally invasive surgical robot, soft tissue localization, visual servoing, tactile sensing, multimodal control

1. Introduction

Robot minimally invasive surgery cuts the skin small, and the patient gets better fast. Visual servoing is a common way to control the robot's position in surgery [1]. In a true surgical procedure though, the tissue is not still. As the patient breathes, as the heart beats and as the surgeon moves the tool, it moves and changes shape. The tissue cannot be simply treated as a fixed point. During heart surgery, the beating heart and breathing circulate the operating field [2]. Just moving the tool to the right pose is not sufficient to keep the tissue moving and bending if it continues to do so [3].

So people start to study how vision and touch can work together [4, 5]. With a camera, the surgeon can see where the tool is and where the target is, and how the tissue looks and bends. With touch sensors, the system can feel the force, the contact, and the local mechanical state after the tool touches tissue. The two kinds of information appear at different times. How to put them both into the control loop is still a problem.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2027.39075

A camera can show what the tissue looks like and where things are, but it cannot tell how hard the tool is pressing. A touch sensor can feel the force, but it cannot see where the target is before contact. Processing the two separately may obscure the link between them when the tool touches the tissue [5]. So how to use vision and touch together, depending on what step the task is at, is a key question for tracking moving soft tissue.

This paper first talks about what vision can see and what touch can feel, and what each one cannot do. Then it divides the control methods into three groups by which signal is the main one: vision-led, touch-led, and fusion at information or control level. At the end, it talks about the problems that are still there, which are time sync, sensor failure, and testing in real clinics.

2. Visual and tactile information in dynamic soft tissue localization

2.1. Markerless visual servoing

Visual servoing means the camera gives the feedback, and the robot moves based on what the camera sees [1]. Knowing the pose, or the relative pose between the robot and the target, is the basic step [6]. Visual servoing is frequently employed to locate target, grasp objects and track the trajectory in robot manipulation.

In image-based visual servoing, the system looks at the difference between the image now and the image it wants, and then sends commands to the robot [7]. For soft tissue surgery, the camera can be put on the end of the robot, which is called eye-in-hand. During the operation, the tissue bends, slips, or gets covered, and the image points move in non-rigid ways. If the conventional IBVS model assumes that the target is rigid, the points may drift, the control may get worse, and in the end the points may disappear from the image.

IBVS does not need to build the whole three-dimensional shape, and it does not ask for very good calibration. This makes it good for small spaces. However, the tissue bends, the tool blocks the view, and the light changes in the body. All these things make the image points hard to track. The primary task of visual control in moving tissue is to maintain the visibility of the points and to correct the target location in the event that it is moved.

2.2. End-effector tactile and force information

In puncture surgery, the camera can only see the surface of the tissue. It cannot feel how hard the needle is pressing. Tactile sensors at the end of the tool can pick up the contact force and the torque between the tool and the tissue, in real time.

These tactile sensors provide feedback to the surgeon about the force they are applying. They may also be useful to identify the position and morphology of the tissue during surgery. As the robot pushes on the tissue it can create a map of the stiffness of various places, and the doctor can determine where to perform surgery [8].

However, a tactile-only method presents a serious limitation. The tactile sensor only works on touch if the tool has touched something. It does not have any knowledge about the target location prior to contact. Thus it is impossible to find the target prior to the commencement of the operation. The force number is only displayed if the tool is already in contact with the tissue. It cannot predict what lies ahead.

Soft tissue as well is not linear, as well as differs from one person to the next. When considering only the number of the force, the stiffness of the tissue can be confusing. Noise and drift can also be found in small medical touch sensors. The estimate may be incorrect if the entire control is touch

DOI: 10.54254/2753-8818/2027.39075

only. Touch alone cannot provide the entire picture. If the tool is a long distance from the target, a camera is still required to get it to target.

2.3. Complementarity of visual and tactile information

Vision and touch see different information and they work at different times. The primary use and limitations of these are displayed in Table 1.

Table 1. Comparison of visual and tactile information in dynamic soft tissue localization

Information source	Main observable information	Suitable role	Main limitations
Markerless vision / endoscopy	Relative position of instrument and target, tissue surface appearance and deformation	Global navigation, target tracking, and pre-contact localization	Occlusion, reflection, weak texture, and non-rigid deformation affect feature stability
End-effector touch / force	Contact events, contact force / torque, and local mechanical changes	Contact protection, local state judgment, and force control	No information before contact; noise, drift, and inter-individual tissue variation
Vision + touch	Global position + local contact state	Simultaneously supports localization and contact correction	Requires synchronization, coordinate / state consistency, and fault handling

3. Visual-tactile cooperative control approaches

3.1. Vision-dominant with tactile compensation

The first type is vision-led with tactile backup. In this case, the vision is primarily responsible for determining the global position. Touch only helps after contact. Vision handles the main position and path control. Touch steps in when the tool touches tissue, when the force gets too high, or when the camera is not sure.

In this setup, the visual servo loop is the main loop. It uses the camera image to navigate, track the target, and find where the tool is before contact. It also plans most of the movement. Touch is simply a side feedback channel. It only joins in after the tool has been placed in contact with the tissue. If the force goes over the safe number, or if the camera loses the target and is not sure about the position, touch kicks in quickly. It corrects the position command from vision and, at the same time, it limits the force to prevent the tool from pressing extremely hard and damaging the tissue.

Early small in-vivo surgical robots mostly focused on seeing, making the tool small, and moving it from far away [9-11]. Later systems started to add force feedback. These studies show that vision and touch can both be used in surgery. But the force feedback mostly goes to the human operator. The robot itself does not use touch to automatically correct for tissue bending.

DOI: 10.54254/2753-8818/2027.39075

3.2. Tactile-dominant with visual correction

Vision cannot control the pressing force very well when the tool is already touching the tissue. Therefore some research takes force and/or touch as the primary feedback. They allow the direct action of the force between tool and tissue to power the robot. In this situation the predominant signal is touch. Vision is only used to locate the target prior to contact or to make slow corrections in position.

This type is more suitable for puncture, palpation or whenever the tool is constantly in contact with tissue. The positive side is that it can directly restrict the strength. Still, it requires the camera to locate the target prior to contact, and to provide the global position if required.

3.3. Information-level fusion and control-level fusion

In the research of robot multimodal sensing, people need to make better prediction about the next step of interaction by using the combination of vision and touching [12]. Vision provides the surgical robot with the global position and the appearance of tissue, touch provides the surgical robot with the local force and whether there is contact with the tissue or not. Together with the other, the system can better cope with situations such as occlusion, drift, or changes of contact, as compared to using one of them only.

There are two input channels. The camera channel takes the endoscopic image, to get the target position and the tissue shape. The touch channel takes the force and torque at the tool end, to get the contact state and how hard it is pressing. The key is not just putting the two numbers together. The two should check each other, at the same task step and at the same time.

The fusion has two layers.

(1) Information-level fusion (perception level): First is the information layer. The camera and the touch sensor each do their own feature extraction and time matching. Then the two channels check each other. This helps reduce mistakes when the image drifts or when one channel loses signal. With a built map of the environment, mobile surgical robots can find their way in a moving soft tissue environment [13].

(2) Control-level fusion (decision level): Second is the control layer. On top of position control, the system adds the touch force, to correct the pose or force command after the tool touches the tissue. They build a network to control position and force together. In this network, spatial reasoning and time sequence in the robot task are consolidated in a unique computing structure to perform different tasks in distinct portions of it. It works pretty well together [14].

4. Challenges and outlook

4.1. Visual-tactile time synchronization and control authority allocation

The camera frame rate, image processing time, touch sampling rate, and the robot control loop speed are not the same. Meanwhile, the tissue keeps moving. If the visual position and tactile force are not simultaneous, merely combining them in one controller will not synchronize them. Moreover, if vision and touch are going to present different information, it is important to know who is "in charge" when, and whether or not the switch will cause the control to jump. The reviewed studies rarely report quantitative results for these problems.

The camera, touch sensor and controller should be given the same time stamp later. They should time how long it takes for the signal to be received, processed, sent and acted upon and then

DOI: 10.54254/2753-8818/2027.39075

synchronize the sight and feel of the signal at the same time. They can implement a state machine or confidence scores in the controller. With a clear target, vision is in control. The weight of the force increases gradually after contact. The camera switches to a backup mode if it fails to locate the target or there is an issue with the touch sensor. They shall report the time sync error, switch time, position error before and after switch, and the force jump in test. These values should be reported rather than only the final accuracy.

4.2. Safe fallback under visual occlusion and sensor anomalies

If a tool is blocking the view, the camera can lose the image for a moment or if it reflects or there is blood. Tactile-sensor drift, saturation, or movement of the contact point can also occur. Two modes do not necessarily make the system safer. The system could not detect the broken sensor, and some improper data may grow up by fusion. Few papers discuss its action if one sensor channel fails, such as to exit, stop or slow down.

Later, people should conduct fault injection tests. They should intentionally obstruct the camera, trigger loss of image points, cause drift or max of the touch sensor as well as delay in communication. After that, they observe if the system reduces the weight of the broken sensor, if the system only uses one mode carefully or if the system only stops and waits. The time they find the fault, the maximum contact force, the maximum position error, and the time at which backup begins should be recorded, and whether the backup is successful or not. These measurements allow safety to be quantified.

4.3. Insufficient clinical validation evidence

Papers utilize various types of test beds: simulation, ex vivo tissue, in vitro tissue testbeds and even human patients. However, these levels cannot be combined. A tissue phantom can demonstrate that the algorithm is functional, but with blood, breathing, tool blockage, sterilization rules and real patient differences, nothing can be proven. Additionally, one real patient case cannot substitute a comprehensive safety test.

A step-wise validation should subsequently be carried out, starting with tissue phantom, dead animals, live animals and eventually real patients. The position error, contact force, and other factors (camera can see or not, what the system does in abnormal cases, and the system delay) should be reported at each step. They also have to consider sterilizing the sensor, miniaturizing it, mounting it on the tool and obtaining ethical approval, prior to testing on a real patient.

5. Conclusion

This paper reviews how vision and touch are controlled together for tracking moving soft tissue. It covers visual localization without markers, touch sensing at the tool end, and three control groups: vision-led, touch-led, and fusion at information or control level. Existing studies show that vision is more suitable for giving global position, target, and instrument-relative state information, and touch is better at showing post-contact local forces and contact events, and the two are not just replacements for each other. The key of cooperative control lies in time synchronization, control authority allocation, and fallback under occlusion or sensor anomalies. At present, studies are very different in tissue models, imaging methods, force sensors, control frequencies, and experimental subjects, and the existing evidence is unevenly distributed from phantoms and ex vivo experiments to clinical cases, so a unified control scheme or clinical safety conclusion that is independent of

DOI: 10.54254/2753-8818/2027.39075

specific conditions cannot be formed yet. Future work should unify the time base of vision and touch, and further make clear the stability and applicable scope of cooperative control through fault injection such as visual occlusion and tactile drift, as well as tiered validation using tissue phantoms, animal experiments, and clinical studies.

References

[1] Hernandez-Barragan, J., Villaseñor, C., Lopez-Franco, C., Arana-Daniel, N., & Gomez-Avila, J. (2024). Image based visual servoing with kinematic singularity avoidance for mobile manipulator. *PeerJ Computer Science, 10*, e2559. https://doi.org/10.7717/peerj-cs.2559

[2] Hu, Q. (2024). Research on motion compensation technology for surgical robots based on deep reinforcement learning. University of Electronic Science and Technology of China. https://doi.org/10.27005/d.cnki.gdzku.2024.005581

[3] Huang, X.-Y., Li, W.-H., Lei, Y.-D., Huang, J.-Y., & Yu, H.-B. (2024). Research advance of surgical robots based on monocular vision tracking and positioning system. *China Journal of Oral and Maxillofacial Surgery, 22*(5), 487-492. https://doi.org/10.19438/j.cjoms.2024.05.013

[4] Friebe, M. (2026). Palpation sensing for robotic-assisted surgery. *Die Orthopädie, 55*, 3-10. https://doi.org/10.1007/s00132-025-04753-1

[5] Yang, J. (2025). Research on robotic grasping state perception and pose optimization based on visual-tactile fusion. Foshan University. https://doi.org/10.27960/d.cnki.gfskj.2025.000538

[6] Li, S., de Wagter, C., & de Croon, G. C. H. E. (2022). Self-supervised monocular multi-robot relative localization with efficient deep neural networks. *Proceedings of the 2022 IEEE International Conference on Robotics and Automation (ICRA)*, 9689–9695. https://doi.org/10.1109/ICRA46639.2022.9812150

[7] Flandin, G., Chaumette, F., & Marchand, E. (2000). Eye-in-hand/eye-to-hand cooperation for visual servoing. *Proceedings of the 2000 IEEE International Conference on Robotics and Automation, 3*, 2741–2746. https://doi.org/10.1109/ROBOT.2000.846442

[8] Yuan, Y. (2020). *Research on tactile sensors for minimally invasive surgical robots and tumor palpation detection technology*. Nanjing University of Aeronautics and Astronautics. https://doi.org/10.27239/d.cnki.gnhhu.2020.001576

[9] Reichenbach, M., Frederick, T., Cubrich, L., Bircher, W., Bills, N., Morien, M., Farritor, S., & Oleynikov, D. (2017). Telesurgery with miniature robots to leverage surgical expertise in distributed expeditionary environments. *Military Medicine, 182*(suppl_1), 316–321. https://doi.org/10.7205/MILMED-D-16-00176

[10] Wortman, T. D., Strabala, K. W., Lehman, A. C., Farritor, S. M., & Oleynikov, D. (2011). Laparoendoscopic single-site surgery using a multi-functional miniature in vivo robot. *The International Journal of Medical Robotics and Computer Assisted Surgery, 7*(1), 17–21. https://doi.org/10.1002/rcs.362

[11] Dolghi, O., Strabala, K. W., Wortman, T. D., Goede, M. R., Farritor, S. M., & Oleynikov, D. (2011). Miniature in vivo robot for laparoendoscopic single-site surgery. *Surgical Endoscopy, 25*(10), 3453–3458. https://doi.org/10.1007/s00464-011-1687-0

[12] Li, L., & Thuruthel, T. G. (2024). Bridging vision and touch: Advancing robotic interaction prediction with self-supervised multimodal learning. *Frontiers in Robotics and AI, 11*, 1407519. https://doi.org/10.3389/frobt.2024.1407519

[13] Chen, S. (2025). Research on map maintenance algorithm for mobile robots based on multi-information fusion. Southwest University of Science and Technology. https://doi.org/10.27415/d.cnki.gxngc.2025.000115

[14] Mao, P., Wu, D., Wang, Y., Sun, J., & Yuan, K. (2026). Dual-arm robot control strategy integrating a visual large model and spatiotemporal decoupling architecture. *Computer Engineering and Applications, 1*–15. Available at: https://link.cnki.net/urlid/11.2127.tp.20260917.1140.007