2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob) | 978-1-6654-5849-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIOROB52689.2022.9925374

Orientation Matters: 6-DoF Autonomous Camera Movement for Video-based Skill Assessment in Robot-Assisted Surgery*

Alaa Eldin Abdelaal[1] *Student Member, IEEE*, Nancy Hong[1], Apeksha Avinash[1], Divya Budihal[2], Maram Sakr[3] *Student Member, IEEE*, Gregory D. Hager[4] *Fellow, IEEE* and Septimiu E. Salcudean[1] *Fellow, IEEE*

***Abstract*— Video-based surgical skill assessment is one of the main assessment methods in robot-assisted surgery (RAS). Expert assessors watch videos of trainees, recorded from endoscopic camera feeds, and evaluate their performance using well-established assessment questionnaires. A major drawback of this method, however, is the high variability in scores between different assessors, mainly due to the limited visual feedback provided by the recorded videos. To solve this problem, we propose a new method for six-degree-of-freedom (6-DoF) autonomous camera movement for RAS, which, unlike previous methods, takes into account both the position and 3D orientation information from structures in the surgical scene. We developed a simulation environment to test our method on the "wire chaser" surgical training task from validated training curricula in RAS. In a study with N = 30 human subjects, we show that our proposed method leads to at least 21% more accurate skill assessment and at least 31% less variability in assessment scores than when using a fixed camera view, or camera movement method based only on position information. Our preliminary work suggests that there are potential benefits to autonomous camera positioning informed by scene orientation, and this can direct designers of automated endoscopes and surgical robotic systems, especially when using chip-on-tip cameras that can be wristed for 6-DoF motion.**

I. INTRODUCTION

Video-based skill assessment has been successfully applied in many domains such as education [1] and sports [2]. The main idea of this approach is to record videos of trainees or performers during their tasks. These videos are then used by experts to assess the skill of the trainee/performer This process is referred to as the monitor-evaluate-modify cycle and its merits has strong evidence from research in psychology [3].

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant), in part by the Canada Foundation for Innovation (infrastructure and operating funds), in part by Intuitive Surgical (equipment donation), in part by the C.A. Laszlo Chair in Biomedical Engineering held by Prof. Salcudean, and in part by the Vanier Canada Graduate Scholarship held by Alaa Eldin Abdelaal.

[1]A. E. Abdelaal, N. Hong, A. Avinash and S. E. Salcudean are with the Electrical and Computer Engineering Department, University of British Columbia, 2332 Main Mall, Vancouver, BC Canada. (email: aabdelaal@ece.ubc.ca)

[2]D. Budihal is with Zipline International, 333 Corey Way, South San Francisco, CA. Her contribution was made when she was with the Electrical and Computer Engineering Department at University of British Columbia.

[3]M. Sakr is with the Mechanical Engineering Department, University of British Columbia, 2054-6250 Applied Science Lane, Vancouver, BC Canada.

[4]G. D. Hager is with Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA.

Surgery is one area where video-based skill assessment has been extensively used [4]. In robot-assisted surgery (RAS), the videos used for this purpose usually come from the endoscope/laparoscope viewing the surgical scene. These videos allow surgeons to monitor and evaluate their own (or their trainees') performance with the goal of assessing their skills and identifying areas of improvement. Many validated tools and questionnaires have been developed to standardize skill assessment based on recorded videos such as the Objective Structured Assessment of Technical Skills (OSATS) [5]. In these questionnaires, assessors give scores to trainees based on several metrics such as how they handle the tissue and if the surgical tools collide with sensitive anatomical structures [6].

The quality of the visual feedback in the recorded videos plays a crucial role in video-based skill assessment. Previous research shows that variability between assessors is a major problem. Gingerich *et al* [7] report that this variability comes from the cognitive limitations of the assessors and from their making unjustified inferences because of the lack of information in the videos. Improving the visual feedback quality in these videos can reduce this variability and improve the video-based skill assessment process.

In this paper, we hypothesize that automating the motion of the endoscope can improve the quality of the visual feedback and hence improve the video-based skill assessment process in RAS. The focus of the growing body of literature on automating the endoscope in RAS was solely on how this can improve the actual performance of the task at hand. While this is an important consideration, it ignores the fact that the camera feeds from these autonomous endoscopes are also used by assessors in video-based skill assessment.

In this preliminary study, we consider the case of having a 6-DoF camera in the surgical scene, in addition to the original endoscope, which has been investigated by many groups. For example, our group designed and tested a "pick-up" camera that can be held by one of the tools of the da Vinci Surgical System [8], [9], [10]. This camera can, hence, move in 6-DoF and provide an additional view of the surgical scene. Moreover, the idea of using additional cameras to provide multiple views of the surgical scene has been tested *in vivo* in laparoscopic cholecystectomy as in [11].

The contributions of this paper are as follows:

- We propose a novel autonomous camera method that can improve video-based skill assessment by tracking

both the position and orientation of objects of interest in the surgical scene. This is unlike previous methods which only consider the position information.
- We implement the above autonomous camera concept in a simulated environment of the da Vinci surgical system where a pickup camera is attached to one of its arms following the concept in [8]. We include a motion planning component to satisfy some practical constraints as the camera moves autonomously.
- We assess the benefits of our proposed autonomous camera method on video-based skill assessment in a user study with 30 subjects. We compare our method against both a stationary camera and a point-based autonomous camera method where the camera moves to make a point of interest at the center of the view, without using any orientation information. In our user study, subjects assessed the performance of a simulated surgical training task in recorded videos under the above three camera methods.

II. RELATED WORK

A. Video-based Skill Assessment in Surgery

Video-based methods are extensively used in minimally invasive surgery (MIS) for training and skill assessment. The effectiveness of these methods has been demonstrated in different surgical settings such as conventional laparoscopy [12] and RAS [13]. The effectiveness of video-based skill assessment methods depends on the quality of the visual feedback provided in the videos.

One approach to improve the visual feedback is to use multiple cameras that view the surgical scene from multiple perspectives. Several groups have explored the feasibility and effectiveness of this approach as in [14] and [15]. In our previous work [10], we evaluated the effectiveness of using two cameras inside the abdomen and showed that this led to a modest improvement in the accuracy of surgical skill assessment in RAS.

In all the above studies, the cameras used to record videos are stationary. In this work, we explore the effectiveness of using an autonomous camera system for video-based surgical skill assessment in a RAS setting.

B. Autonomous Cameras in RAS

Automated camera systems in RAS can be categorized based on the source of information used to automate the camera motion into three main categories. The first one is based on the surgical tools. The second is based on the anatomical structures in the surgical scene and the third one is a combination of the first two.

The majority of the existing work is based on moving the camera according to the motion of one or more surgical tools as in [16], [17]. In contrast, Li *et al* [18] propose a method to automate the camera motion based on anatomical structures in a surgical debridement subtask in a dry lab environment. The combination of tools and anatomical structures information has also been studied in the context of automating the camera motion as in [19].

In all the above work, a single point is chosen and the camera moves to make it at the center of the field of view (FoV), without considering the camera orientation when doing so. There are infinite number of possible camera orientations that can make a point at the center of the FoV. We are not aware of any autonomous camera method that takes the camera orientation into consideration. Our study aims at filling this gap in the literature.

Furthermore, all the above methods have not been tested in the context of video-based skill assessment. Our work also aims at studying the effect of automating the camera on video-based skill assessment in RAS.

III. PROPOSED METHOD

A. Overview

In surgical practice, surgeons often identify several landmarks in the surgical scene and orient the camera with respect to these landmarks. For example, the liver and pancreas should be horizontal in the surgical scene in upper abdominal procedures [20]. In other words, the camera's viewing plane should be parallel to the plane that combines the liver and pancreas. Similar guidelines are used in other procedures such as radical prostatectomy [21]. The correct camera orientation is crucial as it guides the surgeon on where to make the next action in the scene (e.g., where to cut or dissect while avoiding major veins) [22] [23].

Based on the above, the surgical scene can be abstracted as a scene with multiple landmarks/structures which can be combined into a plane of interest. In our proposed method, the camera moves so that its viewing plane will be at a specific orientation with respect to the plane of interest. In addition, the camera position is adjusted to make the landmark(s) of interest at the center of the FoV.

B. Problem Formulation and Motion Pipeline

We assume that we have a plane of interest where: $\boldsymbol{p}_a$ is the center point of the plane and $\mathbf{n}$ is the normal to the plane at $\boldsymbol{p}_a$. The goal is to compute the position and orientation of the camera. The camera target position is $\boldsymbol{p}_t$. The camera's target orientation can be fully described by a frame $\mathcal{F}$ attached to it, as seen in Fig. 1. It is mathematically represented by a rotation matrix $\mathbf{R}_t$ where $col_1[\mathbf{R}_t]$ (the first column in $\mathbf{R}_t$) represents the *x-axis*, $col_2[\mathbf{R}_t]$ the *y-axis*, and $col_3[\mathbf{R}_t]$ the *z-axis* attached to the camera.

We compute the camera target position $\boldsymbol{p}_t$ along this normal vector $\mathbf{n}$, at a fixed distance d_f from $\boldsymbol{p}_a$. To avoid collision with tissue, we consider only the space above the anatomical feature(s) and accordingly consider either $\mathbf{n}$ or $-\mathbf{n}$. At each instant, the camera's positional target is set to be the computed point $\boldsymbol{p}_t$ such that:

$$\boldsymbol{p}_t = \boldsymbol{p}_a \pm d_f \mathbf{n} \tag{1}$$

For the camera's orientation, we first align the camera's optical axis with $\mathbf{n}$ by setting $col_3[\mathbf{R}_t]$ (the *z-axis* of frame $\mathcal{F}$) to **-n** (Eq. 2). This assumes that the direction of $\mathbf{n}$ is above the anatomical feature of interest, otherwise, $col_3[\mathbf{R}_t]$ is set

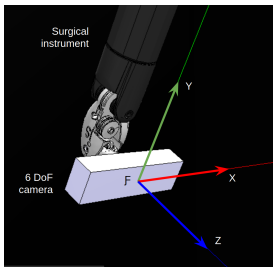

Fig. 1. Frame $\mathcal{F}$ attached to the camera.

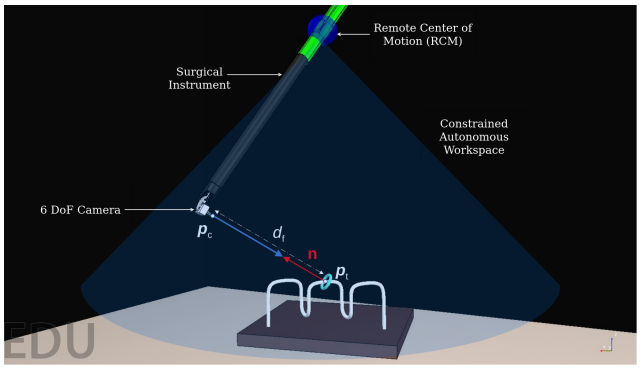

Fig. 2. An annotated setup of our simulated scene showing the camera, feature of interest (cyan ring), constrained workspace, and the wire chaser evaluation task. The wire chaser rail pattern is the same as that in the validated curriculum in [25].

to **n** instead of **-n**. Another desirable characteristic when moving the camera in surgery is to maintain a correct camera horizon [24]. We achieve this by setting $col_1[\mathbf{R}_t]$ (camera's *x-axis*) to be the cross product between a unit vector that is pointing directly upwards with respect to the surgical environment (i.e. the *z-axis* of the world frame denoted by $z\text{-}axis_w$) and the previously computed $col_3[\mathbf{R}_t]$ (Eq. 3). By doing so, we always obtain an *x-axis* that is parallel to the xy-plane of the world frame. The remaining $col_2[\mathbf{R}_t]$ is simply chosen to be orthonormal to the other two axes and this completes the desired rotation matrix $\mathbf{R}_t$ (Eq. 4).

$$col_3[\mathbf{R}_t] = -\mathbf{n} \tag{2}$$

$$col_1[\mathbf{R}_t] = \frac{z\text{-}axis_w \times col_3[\mathbf{R}_t]}{\|z\text{-}axis_w \times col_3[\mathbf{R}_t]\|} \tag{3}$$

$$col_2[\mathbf{R}_t] = \frac{col_3[\mathbf{R}_t] \times col_1[\mathbf{R}_t]}{\|col_3[\mathbf{R}_t] \times col_1[\mathbf{R}_t]\|} \tag{4}$$

Fig. 2 shows an instance of applying the above method to a simulated surgical training task. The plane of interest is the face of the ring and the camera motion is automated to make its viewing plane parallel to the face of the ring while keeping the ring itself at the center of the FoV.

To facilitate collision avoidance, motion planning is incorporated when the distance between consecutive target positions is larger than a set threshold. We generate a set of intermediate waypoints (IWP) between the current camera position $\boldsymbol{p}_c$ and the target camera position $\boldsymbol{p}_t$ as seen in Fig. 3.

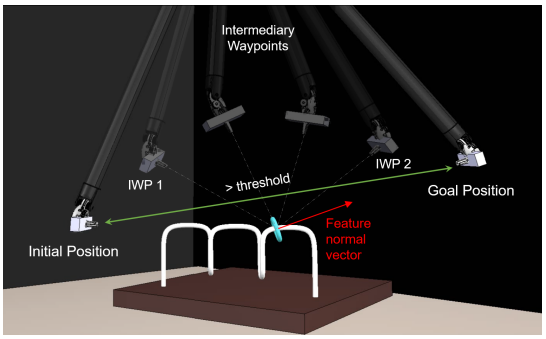

Fig. 3. The motion planning part of our autonomous camera method where IWP are generated.

IWP 1 and 2 are distributed evenly along the line segment between $\boldsymbol{p}_c$ and $\boldsymbol{p}_t$ and are a user-defined distance d_{wp} above the current and target positions, respectively. Using these four points ($\boldsymbol{p}_c$, $\boldsymbol{p}_t$, IWP 1 and IWP 2), a linear function is interpolated and a trajectory is obtained. The orientation of the camera at each of these new intermediate positions is adjusted to ensure that the anatomical feature/structure of interest is always centered in the view. The *z-axis* is set to the vector between the current intermediate point and the feature's center point. The *x-axis* is adjusted to correct for the horizon as described previously, and the *y-axis* is chosen to be a vector orthonormal to both.

Without human-in-the-loop control, it is essential to incorporate safety features into any autonomous system. To this end, we define a constrained workspace (see Fig. 2) within which the camera moves autonomously. Outside this workspace, the autonomous algorithm freezes. We chose to define this constrained workspace in the form of a 3D cone (whose projection appears in blue in Fig. 2), with the following parameters: the cone tip is the remote center of motion (RCM), the cone height is slightly smaller than the length of the surgical instrument/endoscope, and the cone base radius is empirically chosen to be 10 cm. The cone's directional vector is initially set to the vector joining the RCM and the initial position of the feature of interest, to ensure that this feature is always in view, and remains unchanged after initialization.

The proposed autonomous camera method is a Cartesian-based method. This means that it is applicable to any 6-DoF robotic camera regardless of its kinematic configuration. This includes cameras that can be picked up by the normal robot tools as our proposed pick up camera in [8]. This also includes articulated cameras such as the endoscope of the single-port da Vinci System [26].

IV. EXPERIMENTAL EVALUATION

A. Experimental setup

We use the first generation da Vinci system simulator proposed in [27]. It simulates the full patient-side cart of the da Vinci system including two patient-side manipulators (PSMs) and a 4-degree-of-freedom endoscopic camera manipulator (ECM). Controlling the motions of the PSMs

and ECM can be performed in the same way as controlling the patient-side cart in the real robot using the da Vinci Research Kit (dVRK) [28]. The simulator comes with some pre-prepared scenes of different tasks and it also allows adding new scenes/environments as needed. We used this feature to add the wire chaser task scene as shown in Fig. 2, which is described in IV-B.

We modified the simulated da Vinci system by attaching two vision sensors to the end of the surgical tool tip of one of the PSMs to form our 6-DoF camera as shown in Fig. 1. This is similar to the "pick-up" stereoscopic camera concept proposed in our previous work [8].

The use of the simulator allows us to fix all the variables in our evaluation except the autonomous camera method. This enables us to quantify the effect of changing this method on the experiments' outcomes.

For this work, we focus solely on evaluating the effect of our proposed autonomous camera method on video-based surgical skill assessment and not on tissue tracking. That is why we obtain and use ground truth data such as position and normal vector of the structure of interest from the simulator. In a real scenario, this data can be obtained through a dedicated vision pipeline as in [29], [30]. It can also be obtained in image-guided interventions where pre- and intra-operative images are registered, which gives more information about the surgical scene [31].

B. Task

We test our autonomous camera method on the "wire chaser" task which is part of the validated training curricula in RAS [32]. The task has also been validated for multiple surgical specialties such as urology, gynecology and general surgery [33]. Previous research has shown that the level of performance in this task is correlated with the level of performance in the operating room [34].

The task involves holding a ring and moving it along a rail/wire. It is designed to measure the manual dexterity, hand-eye coordination and camera control skills of trainees. Errors in this task occur when the ring touches the rail. The task has also been used in the context of robot-assisted surgical training as in [35] and in video-based surgical skill assessment as in [10].

The ring represents the anatomical structure that we are interested in. The main idea is that a good visualization of the ring (as seen from the camera) is crucial to the wire chaser task. That is why our autonomous camera method uses both the plane of the face of the ring as well as the ring's center as its inputs. The camera then moves autonomously following our proposed method in Section III. That is, the camera moves so that its viewing plane is always parallel to the plane of the face of the ring, and that the center of the ring is always at the center of the FoV.

We argue that results on this task can be generalized to other surgical tasks for multiple reasons. *First*, based on our abstraction of the surgical scene in III-A, the landmarks in a surgical scene are represented in the wire chaser task by the points on the face of the ring. *Second*, the plane of interest in a surgical scene is represented by the plane of the face of the ring. *Third*, the camera viewing plane in the wire chaser task should be parallel to the face of the ring, just like the surgical guidelines in the upper abdominal procedures to do the same with the plane combining the liver and pancreas. *Moreover*, the motion of the ring in the chosen task represents a stress test to our proposed method with different instances of possible orientations of the plane of interest. *Furthermore*, the quality of visually assessing the skill in the wire chaser task is based on how good the camera orientation is. This is exactly like the case in surgical practice when the correct camera orientation guides the surgeon on where to perform the next action. *In addition*, the ability to measure the ground truth errors of the wire chaser tasks in the simulator provides us with a baseline to quantify the benefits of visually detecting these errors based only on the views from the camera.

Our hypothesis is that *using the proposed autonomous camera method, human assessors can better spot the cases when the ring touches the rail compared with other methods that automate the camera based only on position information*. The position information in this case is the position of the center of the ring.

This hypothesis is tested in the context of video-based skill assessment where subjects watch videos of the task to assess the skill using specific criteria. This context is similar to the real context in video-based skill assessment where assessors need to figure out things such as whether or not the trainee touches a critical structure in the scene (e.g., nerves). Such criterion is part of the used assessment questionnaires in video-based skill assessment.

We choose to test our proposed method in video-based skill assessment instead of testing it while subjects perform the actual task in this first study of the proposed autonomous camera method. The reason is that video-based skill assessment is performed purely based on the views of the camera. In contrast, while performing the actual task, a user can have the best camera views to guide his/her motion, but still commits errors due to other factors such as the lack of the required motor skill to navigate the ring along a complex trajectory.

C. Trajectory Randomization

We record videos of the task when we automated the ring motion that is held by one PSM to follow predefined trajectories along the rail. Some of these trajectories are ideal according to the following two conditions: (i) The ring is centered with respect to the rail, and (ii) The plane of the face of the ring is always perpendicular to the rail. Other trajectories were randomized by violating one or two of the above conditions. This in turn introduces a number of collisions between the ring and rail.

The trajectory of the ring's motion along the rail is defined by setting control points evenly spaced from start to finish. A control point's position is defined in (x, y, z) coordinates, and orientation in Tait-Bryan Euler angles (α, β, γ) that together represent a single rotation: $\mathbf{R}_{total} = \mathbf{R}_x(\alpha)\ \mathbf{R}_y(\beta)\ \mathbf{R}_z(\gamma)$.

$\mathbf{R}_z$, $\mathbf{R}_y$ and $\mathbf{R}_x$ represent elemental rotations about the z-, y- and x-axes respectively of the simulation world frame. We automate the ring movement in the simulator to follow the pose of these control points.

To introduce collisions between the ring and rail, randomly generated noise is added to the six variables describing each of the control points. By varying the number of control points and the threshold of noise added to the position and orientation, varying levels of difficulty can be represented in the resultant trajectories. For our tests, we chose trajectories with the following parameters: number of control points: 36 and 71, position noise threshold: 3-3.5 mm, angular noise threshold: 10-30 degrees. A higher number of control points introduces a higher degree of variation in the trajectory, providing more touches/collisions between the ring and rail. This resulted in the video of the trajectory with 71 control points being the most difficult one to assess.

D. Performance Metrics

We tested the proposed autonomous camera method in two aspects. The first one is by measuring the tracking errors as the ring moves along the ideal and randomized trajectories. The second one is by conducting a user study where users watch the recorded videos to count the number of touches between the ring and rail. The goal in this second case is to measure how accurate the users are when using the proposed method in comparison with other methods as explained in Section V below.

In the first evaluation method, we measure the tracking accuracy when the ring is moving along ideal and randomized trajectories according to the following three tracking errors:

- Centering error in the image space: This metric refers to the difference in pixels between the position of the center of the ring on the camera view and the position of the center of the view.
- Centering error in the 3D space: This metric is similar to the first one except that it is the difference in millimeters between the 3D position of the center of the ring and the equivalent 3D position of the center of the FoV.
- Orientation error: This metric refers to the angle between the camera optical axis and the vector **n** that is perpendicular to the plane of the face of the ring.

The above three errors are reported as a function of time along the entire trajectory of the ring. We report them in three cases of using ideal trajectories and in another three cases of using randomized ones.

In the user study, the performance metric is the absolute difference between the reported number of errors (instances of the ring touching the rail) by each participant and the ground truth from the simulator

V. User Study

We conducted a user study (N = 30) to measure the effectiveness of the proposed autonomous camera method while performing a video-based skill assessment task as described above. We recorded videos of the wire chaser task as the ring follows different randomized trajectories under three conditions for the camera motion as follows:

- Condition I is when the camera is fixed, showing the entire task. This represents the baseline condition.
- Condition II is when the camera motion is automated to follow the center point of the ring, regardless of the ring's orientation. This represents an autonomous camera method that is based solely on position information. We refer to this method as the "centering method".
- Condition III is when the proposed autonomous camera method is applied. That is, when the autonomous camera motion is based on both the position and orientation information as described in Section III above.

In the last two conditions, the camera's initial pose was the same as in the fixed camera condition (condition I) above.

This was a within-subject user study, where each subject was exposed to all the study conditions. We recorded a total of nine videos, three per each condition. The nine videos were for the ring moving along the rail in three randomized trajectories with varying levels of difficulty. The goal is to measure the skill assessment accuracy in the videos where the ring moves in the most difficult trajectory (that is, the one with highest level of randomness which is the trajectory that has 71 control points).

Each subject watched the nine videos in three sets, each set containing the three videos of each condition. Subjects were asked to count the number of touches between the ring and rail in each video. Counterbalancing was employed to reduce/eliminate the effect of any learning or carryover bias that may exist when a subject is exposed to each condition. The Latin squares [36] method was used to compute the order in which each subject is exposed to a condition. Since the study has three conditions, we applied two Latin squares, the second one being the mirror of the first, which led to having a total of six cases representing all the six possible combinations of the three conditions.

Due to the restrictions of inviting subjects to the lab (because of the COVID-19 situation), the study was conducted virtually by sending an electronic form to each subject containing the videos. We added an attention question in the middle of the form to make sure that subjects were paying attention. The data of any subject who provided a wrong answer to this question was excluded.

All our subjects had little or no exposure to surgery. Previous research shows that crowd-sourced video-based surgical skill assessment with non-experts is as accurate as the skill assessment performed by expert surgeons and surgical educators [37]. The user study was approved by the Research Ethics Board at the University of British Columbia.

VI. Results and Discussion

A. Tracking Accuracy Results

Based on the performance metrics outlined in IV-D, we conducted two tests with the wire chaser task to evaluate the accuracy of our implemented algorithm. In the first test, the trajectory represents the ideal trajectory of the ring along the

rail, without any collisions between the two. We compute the three metrics: centering error with respect to the left image space, centering error in the 3D space, and orientation error. This test is repeated three times to show the repeatability of our algorithm. Across all three trials, we obtained an average image centering error of 35 ± 37 pixels, 3D centering error of 3 ± 3 mm, and orientation error of 5 ± 12 degrees.

In the second test, we chose three noisy trajectories by adding noise to the ideal positions and orientations of the ring across its path such that the ring collides with the rail at certain points. The three noisy trajectories are the same trajectories used in the user study described in Section V. We obtained an average image centering error of 38 ± 22 pixels, 3D centering error of 3 ± 2 mm, and orientation error of 4 ± 11 degrees across the three paths.

It is important to note here that the tracking accuracy under noisy trajectories is comparable to the case of having ideal ones. This shows the robustness of our proposed method to added noise to the ring's motion, which is the closer case to a human-generated motion for the same task.

Compared with the centering method, our proposed method is at least six times better in terms of the orientation error. At the same time, the centering method is at most 28% better (or 11 pixels more accurate) than our proposed method in terms of the image centering error. This shows that the addition of the 3D orientation tracking did not practically compromise the important need to keep the feature of interest at the center of the FoV.

B. User Study Results

As for the user study, we report the assessment errors in the most difficult video, that is, the one with the highest level of randomization. From the 30 participants in the user study, three participants were excluded after providing a wrong answer to the attention question. We then compared between the subjects' assessment errors across the three study conditions.

As shown in Fig. 4, using the proposed autonomous camera method leads to lower number of assessment errors and less variance between the subjects' scores compared with the other two conditions. In particular, the proposed method (condition III) leads to 25% and 21% fewer assessment errors compared with the centering method (condition II) and fixed camera method (condition I), respectively. Furthermore, the standard deviation in the assessment errors using the proposed method is 31% and 32% lower than that of the centering method and fixed camera method, respectively.

These reductions in the average and standard deviation of the proposed method show its potential to improve the current practice in video-based surgical skill assessment. Previous research in this area show that variability between assessors is a major practical problem due mainly to the cognitive limitations of the assessors. This sometimes leads them to make unjustified inferences which affects the accuracy of their assessment. Our proposed autonomous camera method has the potential to contribute to solving these problems as it can provide better visual feedback which allows the assessors to make more informed assessments and reduces their need to infer/guess due to the lack of the available visual information.

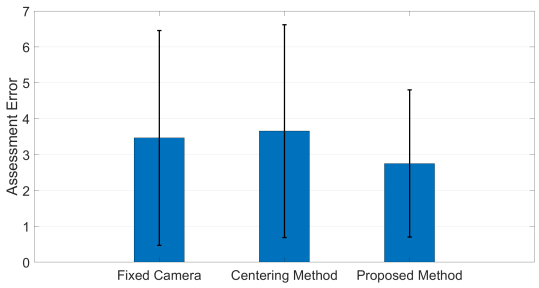

Fig. 4. The results of the user study based on the assessment errors of the subjects in the most difficult skill assessment video.

VII. CONCLUSIONS

Orientation matters in viewing the surgical scene in video-based skill assessment. In this paper, we presented an autonomous camera method for 6-DoF endoscopic camera systems in RAS. Our method takes into consideration both the position and orientation information of anatomical structures of interest in the surgical scene. Our method achieved an average position tracking accuracy of 3 mm and orientation tracking accuracy of 5 degrees when tested on a validated RAS training task in a simulated environment. Moreover, our results show the robustness of our method to added noise to the anatomical structure's motion and that the consideration of 3D orientation did not practically compromise the need to keep the object of interest at the center of the FoV.

We also tested the effectiveness of using our autonomous camera method for video-based surgical skill assessment. We conducted a user study (N = 30) where subjects watched videos of a simulated surgical training task under different camera automation conditions. Our results show that using the proposed autonomous camera method leads to up to 25% more accurate skill assessment and up to 32% lower standard deviations between different assessors in the "wire chaser" task. These results demonstrate the potential of the proposed autonomous camera method in augmenting the cognitive abilities of assessors by providing better visual feedback of the tasks compared with the other methods.

Our results show the importance of including orientation information into automated camera systems in RAS. With the extensive research on articulated endoscopic cameras, the kinematic constraints of tracking such information in practice are removed. Our future work includes improving the proposed autonomous camera pipeline to address potential problems in the visual feedback such as occlusions, and testing the proposed system with subjects conducting a surgical task on a RAS platform such as the da Vinci system.

ACKNOWLEDGMENT

We would like to thank Jordan Liu for his assistance with use of the simulator for this work.

REFERENCES

[1] M. Sherin and E. van Es, "Using video to support teachers' ability to interpret classroom interactions," in *Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), 2002, pp. 2532–2536.

[2] B. D. Wilson, "Development in video technology for coaching," *Sports Technology*, vol. 1, no. 1, pp. 34–40, 2008.

[3] C. S. Carver and M. F. Scheier, *On the Self-Regulation of Behavior*. Cambridge University Press, 2001.

[4] M. G. Goldenberg and T. P. Grantcharov, "Video-analysis for the assessment of practical skill," *Tijdschrift voor Urologie*, vol. 6, no. 8, pp. 128–136, 2016.

[5] J. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown, "Objective structured assessment of technical skill (OSATS) for surgical residents," *British Journal of Surgery*, vol. 84, no. 2, pp. 273–278, 1997.

[6] I. Van Herzeele, R. Aggarwal, I. Malik, P. Gaines, M. Hamady, A. Darzi, N. Cheshire, F. Vermassen, E. V. R. E. R. T. EVEResT, *et al.*, "Validation of video-based skill assessment in carotid artery stenting," *European Journal of Vascular and Endovascular Surgery*, vol. 38, no. 1, pp. 1–9, 2009.

[7] A. Gingerich, J. Kogan, P. Yeates, M. Govaerts, and E. Holmboe, "Seeing the black boxdifferently: assessor cognition from three research perspectives," *Medical Education*, vol. 48, no. 11, pp. 1055–1068, 2014.

[8] A. Avinash, A. E. Abdelaal, P. Mathur, and S. E. Salcudean, "A pickup stereoscopic camera with visual-motor aligned control for the da vinci surgical system: a preliminary study," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1197–1206, 2019.

[9] A. Avinash, A. E. Abdelaal, and S. E. Salcudean, "Evaluation of increasing camera baseline on depth perception in surgical robotics," in *In Proc. of IEEE International Conference on Robotics and Automation (ICRA '20)*, 2020, pp. 5509–5515.

[10] A. E. Abdelaal, A. Avinash, M. Kalia, G. D. Hager, and S. E. Salcudean, "A multi-camera, multi-view system for training and skill assessment for robot-assisted surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, p. 13691377, 2020.

[11] D. Oleynikov, M. Rentschler, A. Hadzialic, J. Dumpert, S. Platt, and S. Farritor, "Miniature robots can assist in laparoscopic cholecystectomy," *Surgical Endoscopy And Other Interventional Techniques*, vol. 19, no. 4, pp. 473–476, 2005.

[12] P. Singh, R. Aggarwal, M. Tahir, P. H. Pucher, and A. Darzi, "A randomized controlled study to evaluate the role of video-based coaching in training laparoscopic skills," *Annals of Surgery*, vol. 261, no. 5, pp. 862–869, 2015.

[13] A. E. Abdelaal, M. Sakr, A. Avinash, S. K. Mohammed, A. K. Bajwa, M. Sahni, S. Hor, S. Fels, and S. E. Salcudean, "Play me back: a unified training platform for robotic and laparoscopic surgery," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 554–561, 2018.

[14] H. Wang, K. Sugand, S. Newman, G. Jones, J. Cobb, and E. Auvinet, "Are multiple views superior to a single view when teaching hip surgery? a single-blinded randomized controlled trial of technical skill acquisition," *PloS One*, vol. 14, no. 1, p. e0209904, 2019.

[15] G. Islam, K. Kahol, B. Li, M. Smith, and V. L. Patel, "Affordable, web-based surgical skill training and evaluation tool," *Journal of Biomedical Informatics*, vol. 59, pp. 102–114, 2016.

[16] S. Eslamian, L. A. Reisner, and A. K. Pandya, "Development and evaluation of an autonomous camera control algorithm on the da vinci surgical system," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 16, no. 2, p. e2036, 2020.

[17] O. Weede, H. Mönnich, B. Müller, and H. Wörn, "An intelligent and autonomous endoscopic guidance system for minimally invasive surgery," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA'11)*, 2011, pp. 5762–5768.

[18] J. J. Ji, S. Krishnan, V. Patel, D. Fer, and K. Goldberg, "Learning 2D surgical camera motion from demonstrations," in *Proc. of IEEE 14th International Conference on Automation Science and Engineering (CASE'18)*, 2018, pp. 35–42.

[19] I. Rivas-Blanco, C. López-Casado, C. J. Pérez-del Pulgar, F. Garcia-Vacas, J. Fraile, and V. F. Muñoz, "Smart cable-driven camera robotic assistant," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 2, pp. 183–196, 2017.

[20] J. Zheng, J. Wang, and Y. Li, "I am your eyes-the reflection of being a camera-holder in laparoscopic gastrointestinal surgery," *Annals of Laparoscopic and Endoscopic Surgery*, vol. 2, 2017.

[21] A. Tewari, J. Peabody, R. Sarle, G. Balakrishnan, A. Hemal, A. Shrivastava, and M. Menon, "Technique of da vinci robot-assisted anatomic radical prostatectomy," *Urology*, vol. 60, no. 4, pp. 569–572, 2002.

[22] A. Pandya, L. A. Reisner, B. King, N. Lucas, A. Composto, M. Klein, and R. D. Ellis, "A review of camera viewpoint automation in robotic and laparoscopic surgery," *Robotics*, vol. 3, no. 3, pp. 310–329, 2014.

[23] P. Cappabianca, L. M. Cavallo, I. Esposito, and M. Tschabitscher, "Transsphenoidal approaches: endoscopic," in *Cranial, Craniofacial and Skull Base Surgery*. Springer, 2010, pp. 197–212.

[24] S. Shetty, L. Panait, J. Baranoski, S. J. Dudrick, R. L. Bell, K. E. Roberts, and A. J. Duffy, "Construct and face validity of a virtual reality–based camera navigation curriculum," *Journal of Surgical Research*, vol. 177, no. 2, pp. 191–195, 2012.

[25] H. Schreuder, C. Van Den Berg, E. Hazebroek, R. Verheijen, and M. Schijven, "Laparoscopic skills training using inexpensive box trainers: which exercises to choose when constructing a validated training course," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 118, no. 13, pp. 1576–1584, 2011.

[26] R. W. Dobbs, W. R. Halgrimson, S. Talamini, H. T. Vigneswaran, J. O. Wilson, and S. Crivellaro, "Single-port robotic surgery: the next generation of minimally invasive urology," *World Journal of Urology*, vol. 38, no. 4, pp. 897–905, 2020.

[27] G. A. Fontanelli, M. Selvaggio, M. Ferro, F. Ficuciello, M. Vendittelli, and B. Siciliano, "A v-rep simulator for the da vinci research kit robotic platform," in *Proc. of the IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob'18)*, 2018, pp. 1056–1061.

[28] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio, "An open-source research kit for the da vinci® surgical system," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA'14)*, 2014, pp. 6434–6439.

[29] M. C. Yip, D. G. Lowe, S. E. Salcudean, R. N. Rohling, and C. Y. Nguan, "Tissue tracking and registration for image-guided surgery," *IEEE Transactions on Medical Imaging*, vol. 31, no. 11, pp. 2169–2182, 2012.

[30] Y. Li, F. Richter, J. Lu, E. K. Funk, R. K. Orosco, J. Zhu, and M. C. Yip, "Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2294–2301, 2020.

[31] G. Samei, K. Tsang, C. Kesch, J. Lobo, S. Hor, O. Mohareri, S. Chang, S. L. Goldenberg, P. C. Black, and S. Salcudean, "A partial augmented reality system with live ultrasound and registered preoperative mri for guiding robot-assisted radical prostatectomy," *Medical Image Analysis*, vol. 60, p. 101588, 2020.

[32] A. P. Stegemann, K. Ahmed, J. R. Syed, S. Rehman, K. Ghani, R. Autorino, M. Sharif, A. Rao, Y. Shi, G. E. Wilding, *et al.*, "Fundamental skills of robotic surgery: a multi-institutional randomized controlled trial for validation of a simulation-based curriculum," *Urology*, vol. 81, no. 4, pp. 767–774, 2013.

[33] T. Alzahrani, R. Haddad, A. Alkhayal, J. Delisle, L. Drudi, W. Gotlieb, S. Fraser, S. Bergman, F. Bladou, S. Andonian, *et al.*, "Validation of the da vinci surgical skill simulator across three surgical disciplines: a pilot study," *Canadian Urological Association Journal*, vol. 7, no. 7-8, p. E520, 2013.

[34] M. A. Aghazadeh, M. A. Mercado, M. M. Pan, B. J. Miles, and A. C. Goh, "Performance of robotic simulated skills tasks is positively associated with clinical robotic surgical performance," *BJU international*, vol. 118, no. 3, pp. 475–481, 2016.

[35] A. Mariani, G. Colaci, T. Da Col, N. Sanna, E. Vendrame, A. Menciassi, and E. De Momi, "An experimental comparison towards autonomous camera navigation to optimize training in robot assisted surgery," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1461–1467, 2020.

[36] I. S. MacKenzie, "Within-subjects vs. between-subjects designs: Which to use?" *Human-Computer Interaction: An Empirical Research Perspective*, vol. 7, p. 2005, 2002.

[37] C. Chen, L. White, T. Kowalewski, R. Aggarwal, C. Lintott, B. Comstock, K. Kuksenok, C. Aragon, D. Holst, and T. Lendvay, "Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance," *Journal of Surgical Research*, vol. 187, no. 1, pp. 65–71, 2014.