Endoscope Navigation and 3D Reconstruction of Oral Cavity by Visual SLAM with Mitigated Data Scarcity

Liang Qiu
National University of Singapore
qiuliang_sjtu@hotmail.com

Hongliang Ren*
National University of Singapore
ren@nus.edu.sg

Abstract

Nowadays computer-assisted surgery (CAS) technologies have been widely used in many aspects of the medical field such as Minimally Invasive Surgery (MIS) or operation focusing on a small surgical site, which has provided significant benefits to patients. However, it is hard for surgeons to determine the accurate poses and surrounding circumstances of the endoscope, due to some restrictions such as narrow field of view (FOV) and misregistration. In this paper, we propose to apply ORBSLAM with a low-cost endoscope to estimate the location of endoscope and create a 3D map for the oral surgery scene, which imposes considerable challenges compared to other human tissue environments, because of the irregular shape, texture-less surface and non-rigid characteristics of the oral cavity. In general, it is very difficult to detect sufficient and effective data for Visual SLAM to realize accurate localization and 3D dense map mainly due to the scarce feature points extracted from tissues and the rare correct matches. In order to reconstruct a denser map for a texture-less oral cavity, laser light markers are used for generating more features, which can mitigate the problem of data scarcity. Besides, we have validated this approach with some experiments on a silicone model of human head. Comparisons between the trajectory/map obtained from ORBSLAM and the ground truth are also provided.

1. Introduction

Following with rapid development of medical technology, incredible advancements have been achieved in the clinical operation. Open surgeries which need to cut the tissue outside to get access directly to the surgical targets can bring great pain and harm to patients such as relatively larger trauma and longer convalescence time, hence Minimally Invasive Surgery (MIS) has gained considerable attention and favor. Undoubtedly, the endoscope plays a key role in such surgeries to allow surgeons to perform examinations or operations in conjunction with other surgical instruments. However, the field of view (FOV) of the endoscope is narrow and limited, which makes it quite difficult to identify the surrounding conditions of the surgical target [1]. Besides, it cannot provide the intuitive depth information and relative position relationship due to the two-dimensional attribute of the endoscope images [2]. Furthermore, the 2D endoscopic video can only display the surface circumstance while lacking the capability to have an insight into the tissue structure beneath the organ surface. All the problems mentioned above lead to scarce useful information that can be provided to surgeons. That is to say, this technique requires a flexible and skillful operation for surgeons who should have rich related experience.

In order to provide more effective auxiliary information for surgeons and reduce the risk of manipulation, the intelligent medical image technology to expand and enhance the endoscopic view contributes tremendously in computer-assisted surgery (CAS) field. Recently, quite a lot of techniques related to endoscopic videos have emerged or under investigation, trying to overcome the intrinsic drawbacks of the endoscope mentioned above, which opens the way for the development of the medical automatic intelligent system.

Monocular Shape-from-Shading (SfS) can reconstruct the 3D structure of tissue surface without much modification to the endoscope [4]. However, it relies on an assumption related to image processing, namely the light source and the endoscope should keep a certain relative pose relationship [5]. Structrue-from-Motion (SfM) is another technique to obtain the 3D structure of the object scene which exploits the image sequence captured at different places by the camera. It has been applied to the endoscope as well considering some constraints related to deformation of the tissue. Nevertheless, SfM method deals with unordered sequences of the images and requires off-line patch optimization, which cannot satisfy some real-time requirement of operation. In addition, there is

This work is supported by the NUSRI China Jiangsu Provincial Grant BK20150386 & BE2016077 and NMRC Bedside & Bench under grant R-397-000-245-511 awarded to Dr. Hongliang Ren.

Liang Qiu and Hongliang Ren(*corresponding author) are with the Department of Biomedical Engineering, National University of Singapore, Singapore 117575, Singapore, and also with NUS (Suzhou) Research Institute (NUSRI), Suzhou, 215123, China.

another method named Visual Simultaneous Location and Mapping (Visual SLAM) which can cope with the real-time surgical navigation challenge and estimate the intra-operative map of the surgical site at the same time [1]. It can provide the surgeons with the immediate feedback including the endoscope location with respect to human tissues and the surrounding 3D map, and help them to make corresponding decisions precisely [6]. However, some challenges remain in this field. One of them is the homogeneous and texture-less tissue surface which is quite different from the man-made environment and hard to extract the features from. This kind of data scarce will have a fatal effect on the performance of Visual SLAM. The reflection property of wet soft tissues will also bring negative effects. Moreover, without robust feature correspondences, it is also impossible to get accurate localization and mapping in feature-based Visual SLAM [8][13]. Another problem existing in a surgical scene is the deformation of the human tissue arising from respiration, nerve impulse or interaction with the medical tools, which does not satisfy the premise of the application of Visual SLAM, that is, rigid environment, when the deformation of the tissue exceeds a limit [7][9]. The mismatched feature points will also cause the failure of Visual SLAM, so filtering all the biased data or outliers is quite important. Another common situation is the occlusion issue caused by the motion of the surgical instruments during the operation [10].

In this paper, we propose to use the ORBSLAM [11][12] with a low-cost endoscope to estimate its location and reconstruct the 3D scene in an oral cavity. As far as we know, this is the first time to apply ORBSLAM into the oral scene. To solve the data scarce problem due to low-texture surface, the laser light markers are used to mitigate data scarcity problem by creating more artificial features which are easily extracted to make correspondence and generate a denser map. In the following, the overall architecture of the system will be introduced and the results obtained from the experiments based on a silicone model of a human head will be presented and analyzed.

2. Related work

Visual SLAM has received wide attention recently because of the distinct advantages that it can provide real-time localization of the endoscope and generate an intra-operative map of the surgical scene at the same time. A monocular Visual SLAM algorithm based on EKF was proposed in the medical application and validated with human in-vivo endoscopic videos, which is non-invasive, convenient, fast, relatively accurate and robust [10]. However, it cannot obtain sufficient data (enough feature points) to create a dense map and the surgical environment is assumed to be rigid. In [9], ORBSLAM was first used in the endoscope tracking and 3D reconstruction, and the experimental object was in-vivo pigs. Semi-dense map of the tissues inside the pigs is generated by a modified matching method and its accuracy is about 3mm~4.5mm compared to computed tomography (CT) scan, while there is no quantitative analysis about the accuracy of the localization of the endoscope. Moreover, whether the algorithm is equally valid has not been tested when the deformation is getting larger. Another paper [14] proposed a quasi-dense reconstruction which is also based on ORBSLAM compared with the semi-dense map created in [9]. It includes two parts for densification. One is feature-based densification which involves both matched and unmatched features. The other is featureless depth propagation using NCC matching algorithm. In order to evaluate the accuracy, the CT model is used as the ground truth when aligning the SLAM reconstruction with the ground truth using best-fitting similarity transform [15]. The Root Mean Square (RMS) error is 4.9mm, which seems not accurate enough. In [16], Visual SLAM was also used to explore the complicated scene to overcome the drawback of the narrow FOV. Poisson Blending was used to promote the visual fidelity. Furthermore, Visual SLAM can also be applied in fetoscopic interventions with a stereoscopic camera mounted at the tip of a continuum robot [17]. EyeSLAM is a SLAM algorithm applied to human retina, which exploits the vessel detection and matching techniques [18].

Compared with all the related work above, we can find that all the techniques are most applied in the interior tissue of organisms, such as the liver, the esophagus and so on. Besides, the reconstruction maps is not accurate enough as shown in [9][14]. However, our application is an oral cavity which is quite different from other tissue surfaces. The problem of the scarce data and biased matches becomes more intractable. Our method is to combine ORBSLAM with artificial laser markers to realize accurate endoscope tracking and 3D denser oral reconstruction.

3. System overview

3.1. Parameter tuning of ORBSLAM

ORBSLAM is one of the best Visual SLAM algorithms at the moment, which can provide relatively robust and accurate tracking and mapping. Besides, it can also tolerate some small deformation of the tissue while applying it to a medical application.

In order to make ORBSLAM performs better in the oral cavity, we need to tune the parameters set up in the original ORBSLAM, whose application is mostly in the large man-made environment, quite different to our application. In order to mitigate data scarcity problem, here we set the maximum number of extracted feature points to 2000, which can help to find more correspondence and generate more map points. Besides, biased matches will be more in texture-less and homogeneous tissue surface, so we decrease the threshold of Hamming distance by a factor 0.95 to reduce the possibility of mismatching.

3.2. System framework

In this part, the system framework will be introduced. As we can see from Figure 1, the laser light will be projected into the oral cavity to produce artificial patterns which are beneficial to feature extraction in the texture-less and homogeneous surface of the oral cavity. The endoscope will be inserted into the mouth at an appropriate angle. While the endoscope is moved slowly to scan the whole oral cavity, the endoscopic video will be obtained for the following processing.

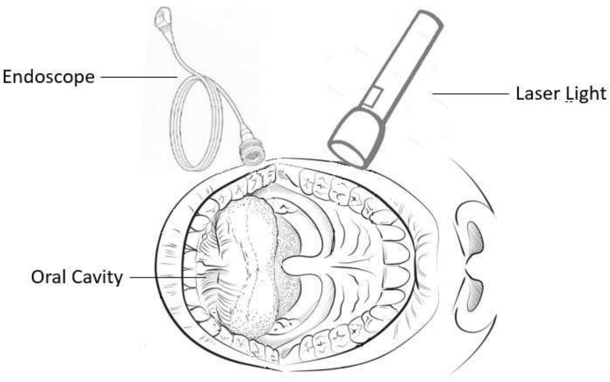


Figure 1. Schematic diagram of the oral SLAM with laser light generating an artificial pattern

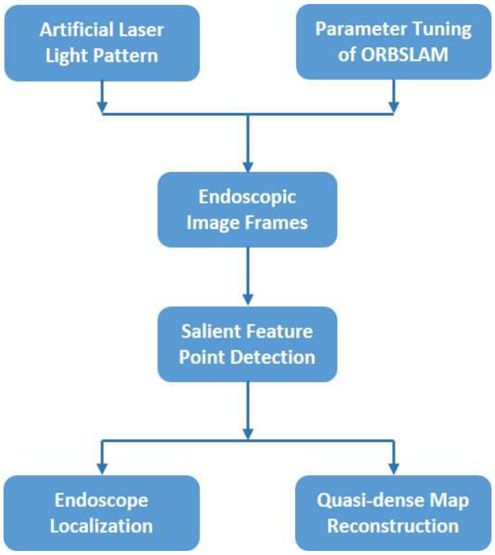

Artificial Laser Light Pattern
Parameter Tuning of ORBSLAM
Endoscopic Image Frames
Salient Feature Point Detection
Endoscope Localization
Quasi-dense Map Reconstruction

Figure 2. The flowchart of our framework

After completing the preparations including laser light setup and parameter tuning of ORBSLAM, the software will try to process the image sequence to get salient feature points. If all things go well, endoscope localization and reconstruction will be realized, shown in Figure 2.

4. Experiment

4.1. System setup

Figure 3 shows the platform of our system. The software is run in Ubuntu 16.04 on an MSI laptop with 7th gen Intel Core i7 processor and 8G RAM. The monocular USB endoscope camera has a white LED whose lightness is adjustable. Its resolution is 640×480 pixels. The Electromagnetic Tracking System (or called EM Tracker) we use is 3DGuidance trakSTAR, which includes an electronics unit, a transmitter, sensors, cables and so on. It uses pulsed DC technology to track the position and orientation of the sensor. Here a sensor is attached to the endoscope to track the trajectory of the endoscope as its ground truth. In order to make the motion of the endoscope more stable and easy to control, a monitor stand is exploited to hold it. The silicone model we use is very close to the real texture and structure characteristics of human. Here the mouth is opened to a certain angle to make it easier to do the experiment. The laser light is fixed right above the mouth of the silicone model with a red holder.

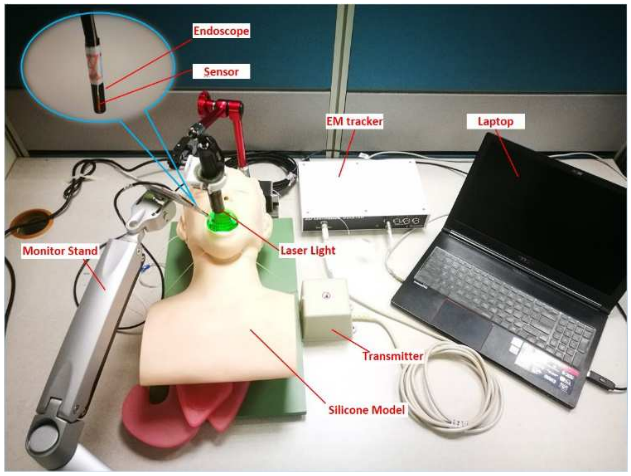


Figure 3. System setup

4.2. Oral SLAM without laser pattern

After tuning the parameters of ORBSLAM, its performance in the oral cavity without the laser pattern will be introduced in this part. In Figure 4 (a-d), we can see that the oral cavity can be divided into several components, such

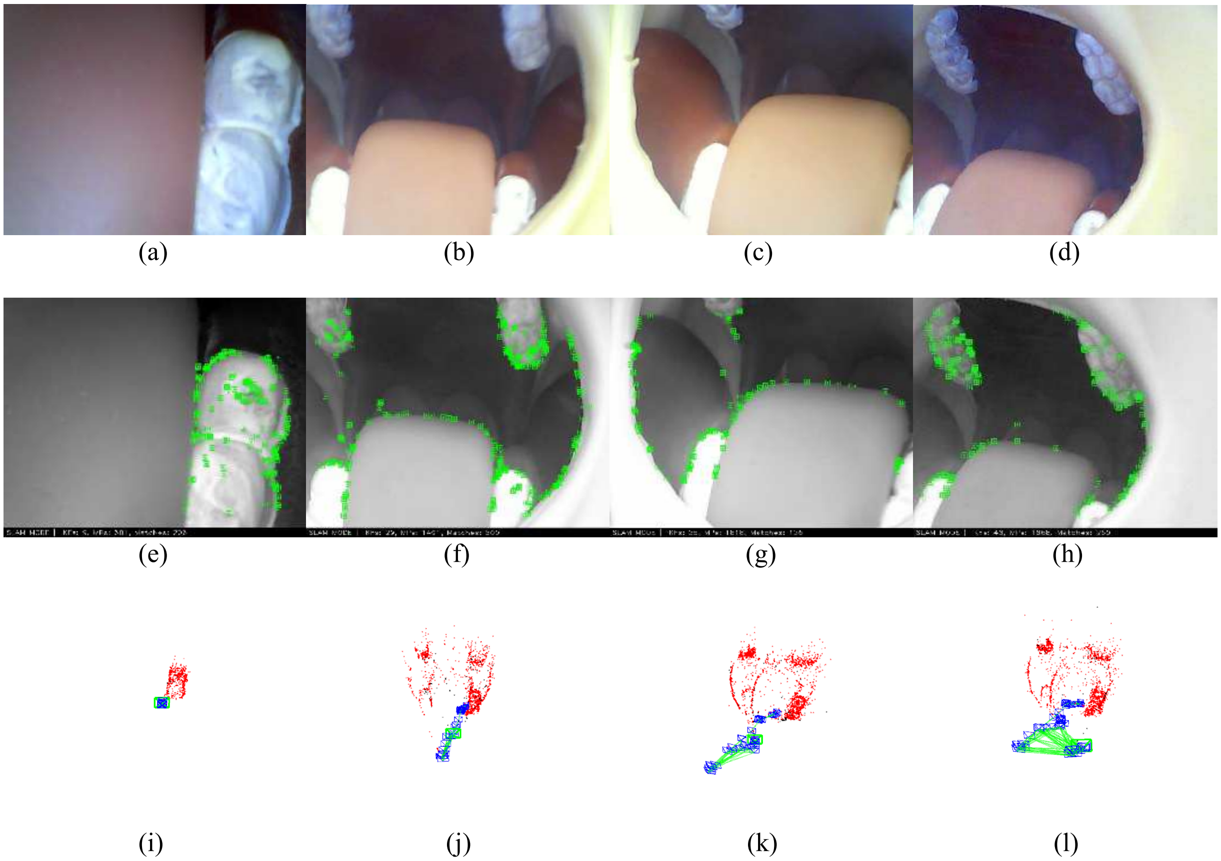


Figure 4. ORBSLAM performance in the oral cavity without laser pattern. (a-d) Original endoscope camera images. (e-h) Gray images with extracted feature points. (i-l) Reconstructed map points and trajectories of the endoscope.

as the teeth, tongue, hard palate, gingivae, and lips. All these parts are soft tissues without obvious textures except the teeth, which poses a great challenge to the ORBSLAM based on discrete feature extraction and correspondence. It is not easy to initialize for ORBSLAM in such scene because of scarce useful feature extraction and matching. Figure 4 presents 4 oral SLAM scenes in chronological order. We succeeded in initializing when the teeth came into the sight of the endoscope camera, shown in Figure 4 (a)(e)(i). Then when the endoscope was moved slowly and stably, more map points were generated and the corresponding Keyframes were also recorded. Finally, the reconstructed map and trajectory of the endoscope are presented in Figure 4 (l).

From the reconstructed maps in each step, we can find most of the map points correspond to teeth and their surrounding area. Other areas of the maps are very sparse due to the homogeneous and texture-less tissue surface. Besides, it should be pointed out that tracking always fails due to lacking of useful feature extraction when the tongue or the hard palate comes into most view of the endoscope.

As we can see, the profile of the oral cavity map is barely visible, which is not friendly interactive information for surgeons. So the denser map which is much more helpful by exploiting more sufficient data must be created. More details of the map such as the number of Keyframes, map points, and matches are shown in Table 1.

Table 1. Map information without laser pattern

Images (Figure 4)	Keyframes (No.)	Map points (No.)	Matches (No.)
(a)(e)(i)	9	581	206
(b)(f)(j)	29	1491	300
(c)(g)(k)	35	1818	136
(d)(h)(l)	43	1868	250

4.3. Oral SLAM with laser patter

From previous experimental results, we can see there are many blank areas or big holes in the generated maps due to scarce feature points in such sites.

In order to reconstruct a denser map, the laser light is used to project laser patterns on the oral surface. By using this method, more feature points can be generated and the

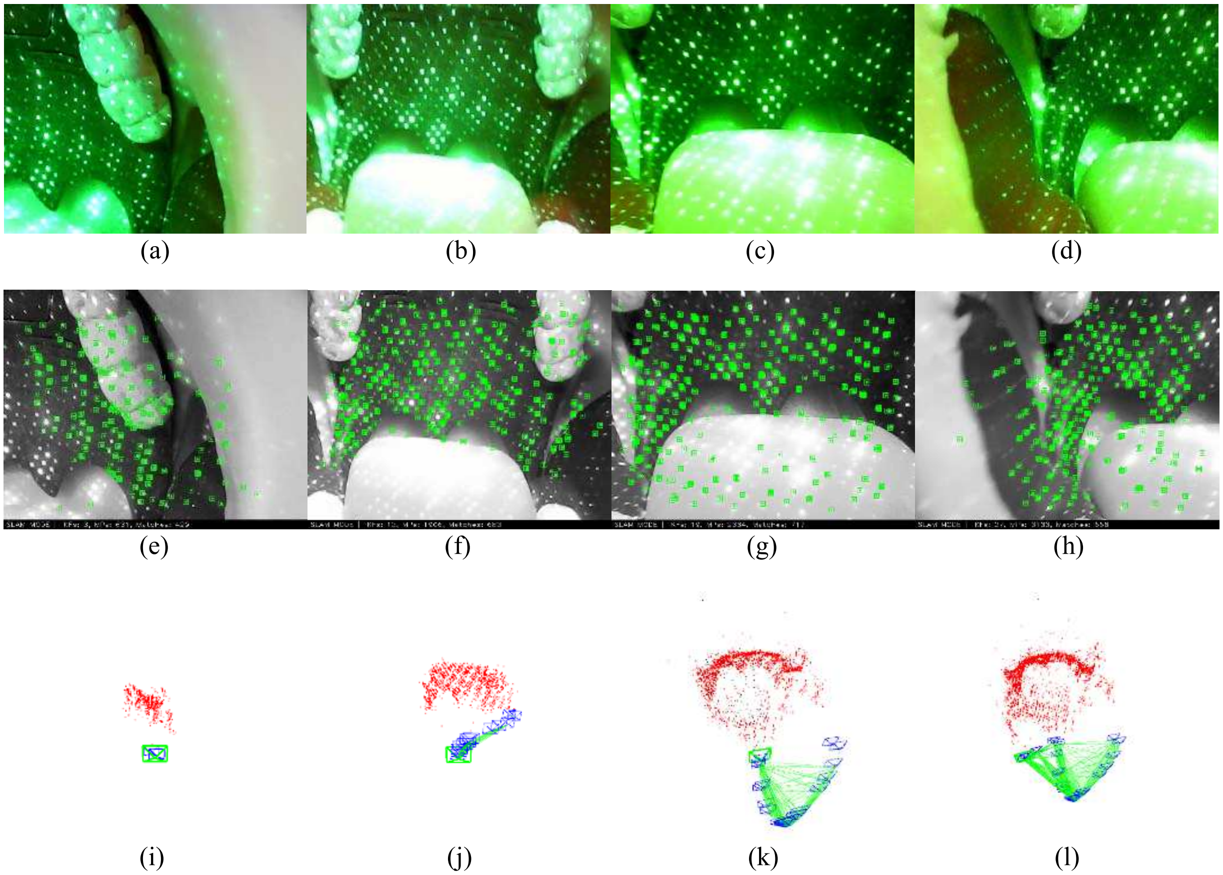


Figure 5. The ORBSLAM performance in the oral cavity with laser pattern. (a-d) Original endoscope camera images with laser patterns. (e-h) Gray images with extracted feature points. (i-l) Reconstructed map points and trajectories of the endoscope.

initialization of ORBSLAM becomes easier and faster, which improves the performance in our application. In Figure 5(a-d), we can see the laser pattern is projected to the surface of the oral cavity model, and feature points can be extracted as shown in Figure 5 (e-h). Notably, more feature points can be extracted and they are well-distributed. The corresponding reconstructed maps and trajectories of the endoscope are displayed in Figure 5 (i-l). From the final reconstructed map shown in Figure 5 (l), we can have a better understanding about the profile of the oral cavity, which can give more auxiliary information for surgeons and do some help to real-time mesh-based denser scene estimation.

Table 2. Map information with laser pattern

Images (Figure 5)	Keyframes (No.)	Map points (No.)	Matches (No.)
(a)(e)(i)	3	631	429
(b)(f)(j)	13	1906	683
(c)(g)(k)	19	2384	717
(d)(h)(l)	27	3133	668

More details of the map such as the number of Keyframes, map points, and matches are shown in Table 2. 3133 map points are generated here, which are much more compared to those (1868 map points) without laser patterns.

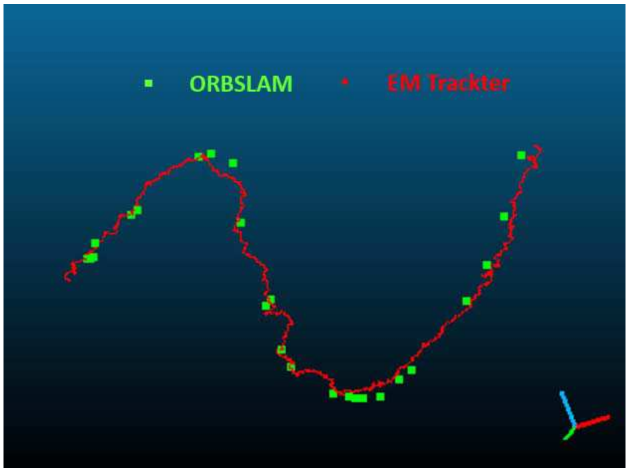
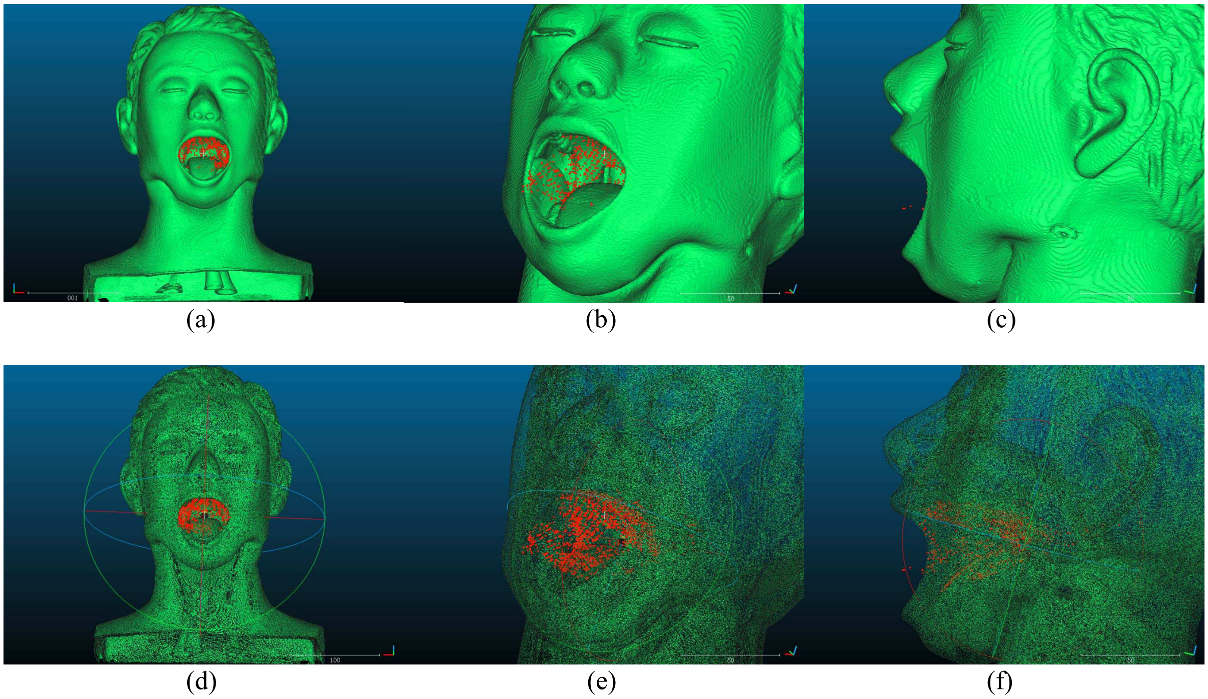


Figure 6. Keyframe positions of ORB-SLAM (green squares) and trajectory ground truth obtained from EM Tracker (red dots)


(a) (b) (c)

(d) (e) (f)

Figure 7. The registration between the CT scan of the sillicone model and the point clound map obtained from ORBSLAM. (a-c) Original 3D CT scan model (green mesh) and reconstructed map (red point cloud) from 3 different viewpoint. (d-f) corresponding semitransparent CT scan with reconstructed map which can provide sharper contrast

Because a sensor of the EM Tracker is attached to the tip of the endoscope, it is easy to get the real-time positions of the endoscope as the ground truth, which is displayed by the red dots in Figure 6. Due to the high sampling frequency of the EM Tracker, all the ground truth data, shown as the red dots, seem to form a curve. However, due to the handheld endoscope which is affected by the unsteady hands, the ground truth curve is not smooth. In our application, we only record the positions of the keyframes, which are discrete points in order to save computing resources and improve the efficiency of ORBSLAM, instead of recording the positions of all the image frames. The Keyframe positions of the endoscope obtained from ORBSLAM are represented by the green squares, shown in Figure 6 as well. In the following, we will try to compare the measured data obtained from ORBSLAM with the ground truth to get the accuracy of our method.

Significantly, the trajectory acquired from ORBSLAM is up to scale because the characteristic of the monocular endoscope, which cannot obtain the actual measured value directly, so if we want to compare the two objects (trajectory from ORBSLAM and the ground truth) with different scales, some registration methods should be exploited. With the estimation of the integrated scale factor, the registration problem can be defined as an optimization problem, shown as formula (1) and (2), according to [19][20].

$$(R,t,s)=\underset{R,t,s}{\arg\min}\sum_{i,j\in D}\|\,gtruth_j - sR\cdot m_i - t\,\|, \quad (1)$$

$$D=\left\{(i,j)\mid gtruth_j\in G,\ m_i\in M\right\}, \quad (2)$$

where G is the set of all the points of the ground truth, while M is the set of all the points obtained from ORBSLAM. Then the RMS error between the tracked positions and the ground truth is 0.765 mm.

Besides, the accuracy of the reconstructed map should also be evaluated compared with the CT scan of the silicone model, using the same registration method mentioned above. In order to improve the registration speed, 6 points which are far away from the oral cavity are removed. This preprocessing will not exert much effects on our analysis because we only focus on the oral part. The final RMS error of the registration is 1.276mm obtained from the remaining 3127 map points.

As shown in Figure 7 (a-c), an original 3D CT scan model (green mesh) and its corresponding reconstructed map (red point cloud) from 3 different viewpoints are aligned. In order to show the distribution of the map points and their relative position relationship compared with the 3D CT scan model clearly, the corresponding semitransparent CT scan with reconstructed map is shown in Figure 7 (d-f).

In order to present a better visualization in terms of the actual values of deviation, the color scale can be used here, where the color saturation range [-1.276, 1.276] is set according to the RMS error, as shown in Figure 8. A more distinct map is shown in Figure 9 by removing the semitransparent CT model.

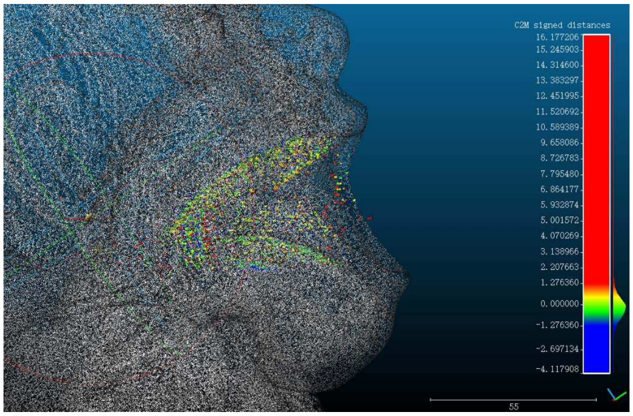


Figure 8. The color scale which can shown the distance compared with the CT reference is applied to the reconstructed map (aligned with semitransparent CT model).

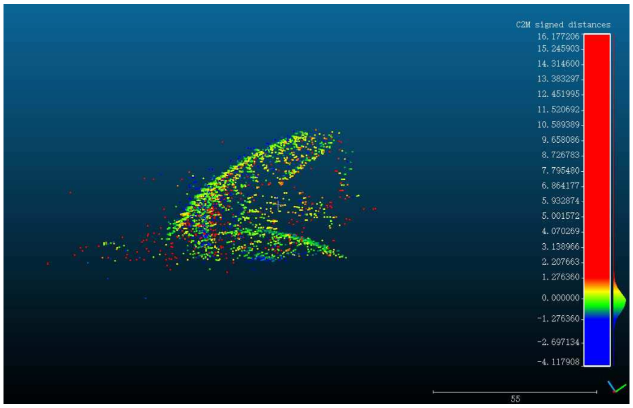


Figure 9. The color scale which can shown the distance compared with the CT reference is applied to the separated reconstructed map.

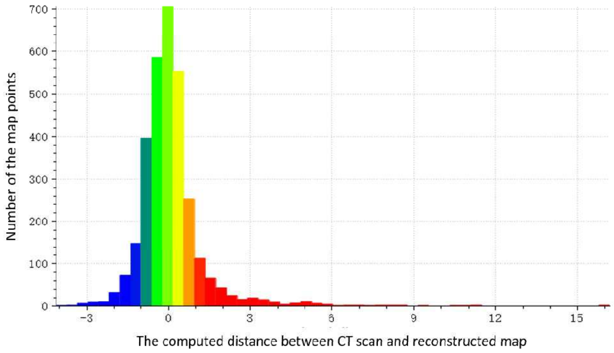


Figure 10. The histogram shows the number of map points belonging to different distance ranges between the CT scan and the reconstructed map.

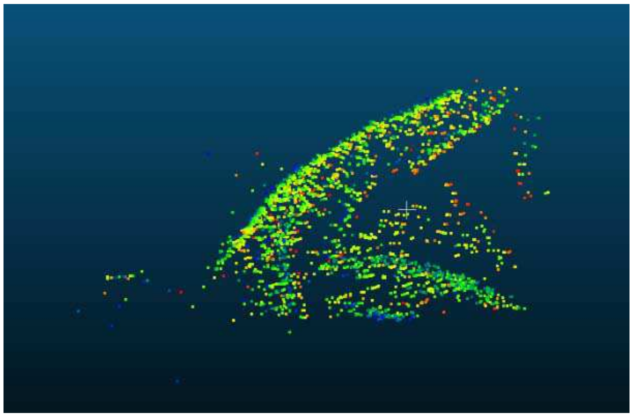


Figure 11. The remaining map after filtered with maximum and minimum threshold according to the computed RMS error (1.276mm) of the distances

Moreover, for better understanding of the map accuracy compared with the CT ground truth, we can see the map point distribution changes for different distance ranges between the CT scan and the reconstructed map in the histogram shown in Figure 10. The remaining map after filtered with the maximum and the minimum thresholds based on the computed RMS error to remove the map points with large errors is shown in Figure 11.

5. Conclusion

In this paper, to realize the accurate localization of the endoscope and the 3D map reconstruction of the oral cavity, we propose to exploit ORBSLAM, one of the best algorithms, with a low-cost endoscope. However, it is very difficult to initialize ORBSLAM and reconstruct a dense map due to the insufficient data obtained from the tissue surface in the oral cavity. Given the challenging scene of the oral cavity which is wet, texture-less and homogeneous, laser patterns are applied to help to generate more feature points and matches to mitigate data scarcity. Besides, the parameters are tuned to acquire more feature points and toughen the standard to filter the mismatches. In this way, the initialization of ORBSLAM will be easier and faster, and a denser map can also be reconstructed compared to the map generated without laser patterns. The experiments have been carried out to demonstrate that the proposed method is feasible in the oral application scenario. The RMS error between the tracked position and the ground truth is 0.765mm, which can meet the needs of most medical applications. Besides, the RMS error for the reconstructed map is 1.276mm, which is relatively accurate to provide more visualization information for surgeons and can be a basis for augmented reality (AR). In the future, non-rigid problems caused by the respiration, the motion of tongue or the interaction with surgical tools in oral cavity will be investigated.

References

[1] Maier-Hein, Lena, et al. "Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery." *Medical image analysis* 17.8 (2013): 974-996.

[2] Bergen, Tobias, and Thomas Wittenberg. "Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods." *IEEE Journal of biomedical and health informatics* 20.1 (2016): 304-321.

[3] Malti, Abed, Adrien Bartoli, and Toby Collins. "Template-based conformal shape-from-motion from registered laparoscopic images." *MIUA*. Vol. 1. No. 2. 2011.

[4] Collins, Toby, and Adrien Bartoli. "Towards live monocular 3D laparoscopy using shading and specularity information." *International Conference on Information Processing in Computer-Assisted Interventions*. Springer, Berlin, Heidelberg, 2012.

[5] Wu, Chenyu, Srinivasa G. Narasimhan, and Branislav Jaramaz. "A multi-image shape-from-shading framework for near-lighting perspective endoscopes." *International Journal of Computer Vision* 86.2-3 (2010): 211-228.

[6] Lin, Bingxiong, et al. "Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey." *The International Journal of Medical Robotics and Computer Assisted Surgery* 12.2 (2016): 158-178.

[7] Mountney, Peter, and Guang-Zhong Yang. "Motion compensated SLAM for image guided surgery." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, Heidelberg, 2010.

[8] Mountney, Peter, et al. "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, Heidelberg, 2006.

[9] Mahmoud, Nader, et al. "ORBSLAM-based endoscope tracking and 3D reconstruction." *International Workshop on Computer-Assisted and Robotic Endoscopy*. Springer, Cham, 2016.

[10] Grasa, Oscar G., et al. "Visual SLAM for handheld monocular endoscope." *IEEE transactions on medical imaging* 33.1 (2014): 135-146.

[11] Mur-Artal, Raul, Jose Maria Martinez Montiel, and Juan D. Tardos. "ORB-SLAM: a versatile and accurate monocular SLAM system." *IEEE Transactions on Robotics* 31.5 (2015): 1147-1163.

[12] Mur-Artal, Raul, and Juan D. Tardós. "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 33.5 (2017): 1255-1262.

[13] Puerto-Souza, Gustavo A., and Gian-Luca Mariottini. "A fast and accurate feature-matching algorithm for minimally-invasive endoscopic images." *IEEE transactions on medical imaging* 32.7 (2013): 1201-1214.

[14] Mahmoud, Nader, et al. "SLAM based Quasi Dense Reconstruction For Minimally Invasive Surgery Scenes." *arXiv preprint arXiv:1705.09107* (2017).

[15] Horn, Berthold KP. "Closed-form solution of absolute orientation using unit quaternions." *JOSA A* 4.4 (1987): 629-642.

[16] Mountney, Peter, and Guang-Zhong Yang. "Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping." *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009.

[17] Dwyer, George, et al. "A continuum robot and control interface for surgical assist in fetoscopic interventions." *IEEE robotics and automation letters* 2.3 (2017): 1656-1663.

[18] Braun, Daniel, et al. "EyeSLAM: Real‐time simultaneous localization and mapping of retinal vessels during intraocular microsurgery." *The International Journal of Medical Robotics and Computer Assisted Surgery* 14.1 (2018).\

[19] Besl, Paul J., and Neil D. McKay. "Method for registration of 3-D shapes." *Sensor Fusion IV: Control Paradigms and Data Structures*. Vol. 1611. International Society for Optics and Photonics, 1992.

[20] Zinßer, Timo, Jochen Schmidt, and Heinrich Niemann. "Point set registration with integrated scale estimation." *International Conference on Pattern Recognition and Image Processing*. 2005.