Article:
Jia, T, Taylor, ZA and Chen, X (2021) Long term and robust 6DoF motion tracking for highly dynamic stereo endoscopy videos. Computerized Medical Imaging and Graphics, 94. 101995. ISSN 0895-6111

https://doi.org/10.1016/j.compmedimag.2021.101995

Long Term and Robust 6DoF Motion Tracking for Highly Dynamic Stereo Endoscopy Videos

Tingting Jia[a], Zeike A. Taylor[b], Xiaojun Chen[c,*]

[a] *School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China*
[b] *CISTIB Centre for Computational Imaging and Simulation Technologies in Biomedicine, Institute of Medical and Biological Engineering, University of Leeds, Leeds, UK*
[c] *School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China*

Abstract

Real-time augmented reality (AR) for minimally invasive surgery without extra tracking devices is a valuable yet challenging task, especially considering dynamic surgery environments. Multiple different motions between target organs are induced by respiration, cardiac motion or operative tools, and often must be characterized by a moving, manually positioned endoscope. Therefore, a 6DoF motion tracking method that takes advantage of the latest 2D target tracking methods and non-linear pose optimization and tracking loss retrieval in SLAM technologies is proposed and can be embedded into such an AR system. Specifically, the SiamMask deep learning-based target tracking method is incorporated to roughly exclude motion distractions and enable frame matching. This algorithm's light computation cost makes it possible for the proposed method to run in real-time. A global map and a set of keyframes as in ORB-SLAM are maintained for pose optimization and tracking loss retrieval. The stereo matching and frame matching methods are improved and a new strategy to select reference frames is introduced to make the first-time motion estimation of every arriving frame as accurate as possible. Experiments on both a clinical laparoscopic partial nephrectomy dataset and an ex-vivo porcine kidney dataset are conducted. The results show that the proposed method gives a more robust and accurate performance compared with ORB-SLAM2 in the presence of motion distractions or motion blur; however, heavy smoke still remains a big factor that reduces the tracking accuracy.

Keywords: Augmented Reality, Minimally Invasive Surgery, 6DoF Motion Tracking.

*Corresponding author

Email addresses: jia2ting@sjtu.edu.cn (Tingting Jia), Z.Taylor@leeds.ac.uk (Zeike A. Taylor), xiaojunchen@sjtu.edu.cn (Xiaojun Chen)

1. Introduction

Minimally invasive surgery (MIS) in the abdomen, in which both visual sensors such as a binocular or monocular laparoscope and surgical instruments are introduced via small incisions, is gaining prevalence due to its reduced trauma compared with open surgery (Bernhardt et al., 2017). However, it introduces new difficulties associated with its limited field of view, lack of inner structure, and lack of force feedback. Video see-through augmented reality (VST-AR) (Nicolau et al., 2011) is a straight-forward idea which aims to supplement intraoperative endoscopy videos with visualizations of blood vessels, tumor, and other sub-surface structures based on preoperative computed tomography (CT) or magnetic resonance (MR) images, to reduce operation difficulties and make surgery more safe, accurate and efficient. Though plenty of VST-AR systems use extra devices such as optical or magnetic tracking system to obtain the motion of the target, the intraoperative videos themselves provide the possibilities for motion tracking.However, producing accurate and robust VST-AR systems remains challenging in dynamic surgical environments. Firstly, target organs move due to respiration, cardiac motion or interaction with surgical tools, and their motion usually differs from that of surrounding tissues because of their different biomechanical characteristics. Secondly, motion of the surgical tools within the field of view themselves bring more movement to the scene, and these may also occlude the target organs. These factors lead to disordered motion fields between neighboring frames. Image motion blur caused by quick retraction/reinsertion of the laparoscope, illumination changes, and specular reflections introduced by light sources bring further difficulties. If the motions of all parts of the target are consistent, only a 6DoF rigid transformation is required to register the preoperative CT model with surgical scene. Otherwise, deformation must be considered.

In this paper, we focus on 6DoF motion tracking for dynamic stereo endoscope videos, and aim to utilize it in clinical VST-AR applications where the deformation of target tissue is negligible. Our target clinical application is laparoscopic partial nephrectomy. Herein, locating the boundary between tumor and normal kidney tissue at the beginning of the procedure is important, because once the smooth membrane between those two is accurately located, it is simple to remove the tumor from the kidney just following this membrane. Presently, this boundary can only be estimated by comparing the model reconstructed from preoperative CT images with intraoperative video images, which is mentally strenuous and difficult to perform accurately, especially when the tumor is inside the kidney. In this process, the kidney's deformation from interaction with surgical instruments during the tumor localization process is negligible and such a VST-AR system can heavily reduce the difficulty in locating the tumor for surgeons. Accurate estimation even of only rigid motion (i.e. 6DoF pose) of the target organ in this way is difficult because of the highly dynamic surgical environment as mentioned above.

Though point clouds registration-based methods and template-based methods can deduce target motion directly, simultaneous localization and mapping

(SLAM) technologies are getting more attention because of their robustness (Chen et al., 2017). SLAM methods can simultaneously estimate the camera's pose and build a map for the captured scene. Strategies for dealing with tracking loss, and local and global pose optimization by bundle adjustment (BA) are also available. Moreover, it is reported that BA-based pose estimation is more accurate than ICP methods on pose optimization (Mur-Artal et al., 2015). Several descriptions of VST-AR applications using ORB-SLAM directly for tracking have been reported (Song et al., 2018b; Mahmoud et al., 2019, 2017b), however, accuracy was limited because of the scene rigidity assumption of SLAM.

Unlike the previous work, we found that accuracy loss due to scene dynamicity of SLAM-based tracking methods for VST-AR applications can be avoided if the motion inconsistencies caused by surgical tools or surrounding tissues are excluded. Herein, the proposed 6DoF motion tracking method takes advantage of the latest 2D target tracking methods and non-linear pose optimization and tracking loss retrieval in SLAM technologies. The SiamMask deep learning-based target tracking and segmentation method is incorporated to roughly exclude motion distraction from surgery tools or surrounding tissues. The computation cost is light compared, for instance, with segmentation approaches like mask R-CNN, which makes it possible for our tracking method to run in real-time, which is critical for our clinical application. In contrast to the frame-by-frame tracking methods, an ORB-SLAM-like global map and set of keyframes are maintained for pose optimization and tracking loss retrieval. The stereo matching and frame matching methods are improved and a new strategy to select reference frame is introduced to make the first time motion estimation of every arriving frame as accurate as possible. Loop closing is one component in traditional SLAM technologies, which detects the images taken by the same camera pose and is regarded as a signal for optimization of motion trajectory. However, few loop closures could be detected on our experiments, mainly due to the small endoscope field-of-view. As a result, loop closure is not part of our system, which is more computationally efficient.

We evaluate the proposed system quantitatively on an ex-vivo porcine kidney dataset, where the gold standard of target organ motion is acquired by an electromagnetic tracking system. Qualitative results are also obtained from experiments with a laparoscopic partial nephrectomy dataset to illustrate its effectiveness further.

2. Related Work

Intraoperative motion tracking methods based on optical or magnetic tracking devices require artificial markers to be affixed to patient skin (Feuerstein et al., 2008; Kong et al., 2017). Such tracking devices are usually expensive and need calibration before surgery. For these reasons, pure visual tracking methods which only utilize images or videos during surgery without any artificial markers are attractive. Markerless 6DoF motion tracking methods can be divided into three categories, that is, point cloud registration-based, template-based and visual SLAM-based methods.

In point cloud registration-based methods, 3D shapes from the surgery scene at different frames, presented as dense or sparse sets of 3D points, are first reconstructed from images using stereo reconstruction, shape from shading, or shape from structure lighting (Lin et al., 2016). A point cloud registration algorithm such as iterative closest points (ICP) (Besl and McKay, 1992) is then used to estimate the relative pose between 3D shapes. Point correspondences are often inferred by some 2D image feature tracking methods, for instance LK optical flow (Allan et al., 2015; Plantefève et al., 2016; Allan et al., 2018)(see also (Bouguet, 2001)) or tracking by feature detection and matching (Kim et al., 2012) to reduce computation costs. In Puerto-Souza et al. (2014) a feature tracking method which is robust to illumination change was used. Additional state estimation frameworks, for instance Kalman filtering or particle filtering, sometimes are introduced to smooth the motion estimate. The authors claimed that feature loss resulting from deformation, motion blur, or occlusion is inevitable, regardless of the feature tracking method used. In response, they proposed a feature matching based tracking retrieval strategy. Feature point based tracking methods also always suffer heavily from accumulated error.

Template-based motion tracking approaches formulate 2D target segmentation and 6DoF motion tracking in a uniform framework. A pre-obtained 3D model is projected onto the 2D image, then the projected silhouette is used as a constraint for a level-set based target segmentation in the 2D image (Prisacariu, 2011). This approach is sensitive to image noise, and may fail if the contour of the target in the image is ambiguous. Allan et al. (2014, 2018) used sparse and dense optical flow respectively with this region-based method for a robust 6DoF tracking of surgical instruments, but were unable to realize a real-time tracking. Wang et al. (2018) proposed combined dense cues with this region-based approach, achieving a real-time performance with computation acceleration on a GPU. However, the performance of such methods rely heavily on the selection of the level-set function and inner parameters configuration of optimization, which may need a cumbersome tuning for different models and scenes. Besides, this method is prone to converge at a local minimum, especially when the contour of the target image is insufficiently clear, or target movement is too fast. These first two classes of methods focus on motion estimation between pairs of frames, which is then repeated frame-by-frame.

Visual SLAM techniques can simultaneously locate a camera's position and build a map of the captured scene in real-time. In contrast to the foregoing methods, they also can relieve accumulated error and relocate after tracking failure. The most representative approach is ORB-SLAM (Mur-Artal et al., 2015) and its stereo and RGB-D version in ORB-SLAM2 (Mur-Artal and Tardos, 2017), however, its rigid scene assumption hinders its direct application to medical VST-AR systems. Some researchers have explored its feasibility nonetheless. In Mahmoud et al. (2017a), ORB-SLAM was applied to tracking patient's position by regarding the surgery scene as static. Similar approaches can be found in Chen et al. (2018); Qiu and Ren (2018); Mahmoud et al. (2019), but the latter works pay more attention to reconstructing a denser map. As the assumption of rigidity of the scene as a whole, as distinct from rigidity of individual struc-

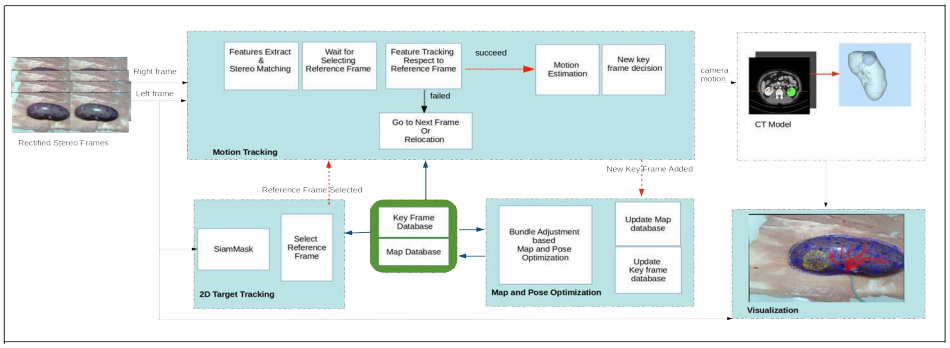

Figure 1: Overview of the proposed system, which contains four threads, namely, 2D target tracking, motion tracking, map and pose optimization and visualization. The former three share a global database which saves map points and keyframes, but only map and pose optimization has both read and write permissions for this; other two have only read permissions. All steps performed by each thread are also shown.

tures within the scene, is likely to be inaccurate in clinical applications, Song et al. (2018a,b) have treated dynamic parts as outliers, and excluded them by a random sample consensus (RANSAC) based iterative pose estimation process. This strategy still cannot guarantee an accurate motion estimation especially in highly dynamic surgery scenes. Identifying dynamic and static parts of images may be helpful in generalizing SLAM techniques to dynamic scenes. Recently, Bescos et al. (2018) proposed detecting dynamic objects by combining the deep learning based object segmentation method Mask R-CNN (He et al., 2017) with multi-view geometry, and demonstrated promising results for SLAM in dynamic environments. Unfortunately, this approach cannot run at real-time speeds due to the high computation cost of the selected object segmentation algorithm.

Object segmentation in video frames has more constraints than segmentation of a single image since adjacent frames have many similarities. Wang et al. (2019) thus proposed a deep learning based approach, named SiamMask. This method can track and segment a selected target simultaneously, and achieves an average frame rate of 55 fps. This technique presents new possibilities for the 6DoF motion tracking field. In this paper, we combine this deep learning based target tracking method with traditional feature tracking. At the same time, we introduce a keyframe management strategy from SLAM techniques to produce a robust yet real-time 6DoF motion tracking method for highly dynamic surgery environments depending only on stereo endoscope videos.

3. Proposed Method

To maximize compatibility with the clinical environment, the proposed 6DoF motion tracking method is designed to be robust to motion distraction from surrounding tissues or surgical tools, able to run at real-time speeds, and easily embedded in a VST-AR system. By choosing the camera coordinate system

of the first intraoperative frame as world coordinate system (WCS), the rigid transformation $\mathbf{T_{wm}}$ mapping preoperative CT model to WCS can be obtained by registering the CT model with target 3D shape reconstructed from the initial frame, which is described in 3.1. For any intraoperative frame k, if the camera motion relative to WCS $\mathbf{T}_k$ is solved, the CT model can be transferred to the camera coordinate system of frame k by $\mathbf{T}_k\mathbf{T_{wm}}$, so that the CT model can be projected to frame k and interior structures from preoperative images can be fused. The proposed 6DoF motion tracking method is aimed at solving $\mathbf{T}_k$. As a feature point based 6DoF motion tracking method, it needs to be capable of feature points detection and matching, motion estimation and optimization, and elimination of additional motion interference. Rapid feature relocation is also essential since tracking loss is inevitable in the presence of motion blur and smoke. Therefore, the proposed method consists of four threads, namely: 2D target tracking, motion tracking, local mapping, and visualization. An overview of the proposed system is presented in Fig. 1.

The raw stereo images are first undistorted and rectified using camera parameters obtained by camera calibration. SiamMask is then used in initial and relocation frames to predict a mask for the selected target which will enable exclusion of other motion distraction in the scene. In the motion tracking thread feature points on both stereo images are detected and matched, and combined with the 2D target tracking result to infer the target's 6DoF motion in the current frame with respect to the selected reference frame. Relocation, where needed, is also performed in this thread. The local mapping thread maintains a set of map points and keyframes, and optimizes these points and pose of keyframes using BA. For each current frame, a reference frame is selected from these keyframes so as to minimize the cumulative error. Once the motion of the current frame is obtained, preoperative CT models are projected into this frame in the visualization thread.

3.1. Preprocessing of preoperative data and oblique-viewing laparoscope calibration

This subsection describes the preliminary steps required by our system, which are CT model reconstruction and laparoscope calibration. Kidney and tumor are segmented manually from their CT dataset by a medical expert using ITK-SNAP (Yushkevich et al., 2006). The corresponding surface models are then reconstructed. Those reconstructed models are registered to the 3D shape represented by a set of sparse points reconstructed from the first frame by an affine-to-coarse point clouds registration method. The three major axes are derived by principal component analysis associated with their centroids for a coarse rigid registration. Then an ICP method is used to achieve a more accurate rigid registration. As this point clouds registration method is not sufficiently accurate, a further manual adjustment through our graphical user interface is performed. We treat an automatic and accurate point clouds registration method as one of our future objectives to make the initialization process fully automatic.

Both intrinsic and extrinsic parameters of a laparoscope are indispensable in stereo reconstruction, however, its calibration process is somewhat different from that of a normal camera because of its special physical structure. The most commonly used binocular laparoscopes are oblique-viewing, which means the scope cylinder axis is tilted from the camera viewing direction, so that the imaging plane is not perpendicular to optical axis Z_c of the mounted scope (see Fig. 2). Therefore, it cannot be modelled as a standard pinhole camera. Fortunately, we find this only stretches the image in the y_i direction of the image coordinates. The lengths y_{is} and y_{cs} of corresponding line segments (see red and green line segments in Fig. 2) aligned with the y-axes in the image and camera coordinate systems, S_i and S_c, are related via the viewing direction angle θ as:

$$y_{cs} = y_{is} \sin\theta \tag{1}$$

Thus, simple oblique-viewing rectification by resizing captured images in the y-direction to a scale using image interpolation will render the optical system equivalent to a standard pinhole camera. It can then be calibrated using generic camera calibration methods, such as Zhang's method Zhang (2000) implemented in Matlab or OpenCV.

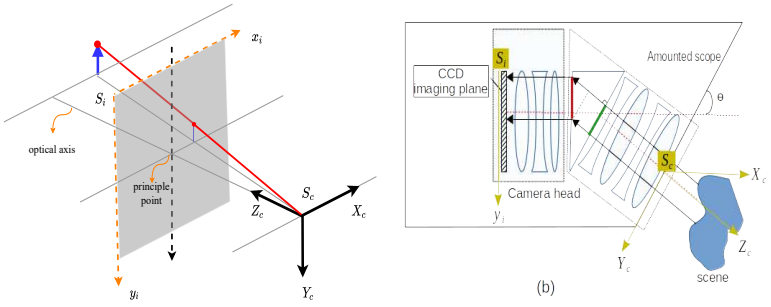

Figure 2: Schematic of two camera models: (a) the standard pinhole camera model, where axes X_c,Y_c of camera coordinate system S_c are parallel to axes x_i and y_i of the image coordinate system S_i; (b) the oblique-viewing camera model, in which the red solid line segment denotes distance of two points in the captured image along the y_i direction, while the green line is the size it should be in a standard pinhole camera model. Fig. 2(b) was modified from (Snaauw (2017)).

3.2. 2D Target Tracking with SiamMask

Region location of the target is required in our proposed method at initial frame and keyframes to exclude motion distraction from the surrounding environment. We use SiamMask, as presented by (Wang et al., 2019), which takes a selected initial bounding box of the target as input, and outputs a binary segmented mask, a similarity score, and a bounding box for searching the image. Essentially, the video target tracking problem is modelled as a template-based class-agnostic object segmentation task. It adopts a popular Siamese-based object tracking approach and adds a binary segmentation task to its loss function.

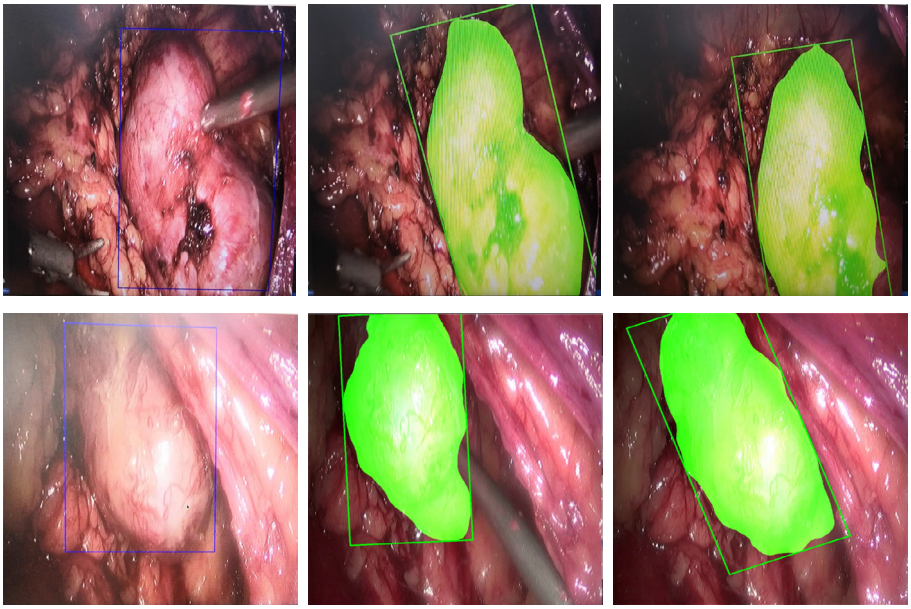

Figure 3: SiamMask 2D target tracking on two clinical surgery datasets (top and bottom rows, respectively). Left images in each row are initial frames, where bounding boxes of targets are manually selected. Middle and right images show targets tracking results in later frames. Both tracking boxes and masks are shown.

Compared with other video instance segmentation methods, such as Mask R-CNN, its biggest advantage is that it can run at about 40 fps on a GTX 1060 GPU with 3 GB memory, which makes it possible for our system to achieve real-time performance.

Our experiment using a pre-trained model to track the kidney in laparoscopic video shows its ability to predict a coarse location for the target. The model was trained by the authors of SiamMask on three public datasets: YouTube-VOS (Xu et al., 2018), COCO (Lin et al., 2014), and ImageNet-VID (Russakovsky et al., 2015). Please refer to Fig. 3 for examples of tracking results.

3.3. Motion Tracking

In this section, the steps for 6DoF motion tracking, which retrieve the motion of the camera relative to the WCS for each frame, are described. Though both laparoscope and captured scene are, in general, in motion, it is still valid for us to approach the tracking by regarding the target as static and camera as moving, since the target is assumed to be without deformation and other motion distractions are excluded.

3.3.1. Feature Points Detection and Matching

Feature points detection and matching plays an important role in both motion estimation and relocation. To detect enough feature points and ensure

a reasonably homogenous distribution, we utilized the same feature detection solution as ORB-SLAM (Mur-Artal et al., 2015), where FAST corners of eight-level image pyramid of each input image are detected, and detector threshold is adopted to try to extract at least five corners per sub-grid of each image level. We compute the ORB descriptor as the feature descriptor for every obtained FAST corner. Feature matching between stereo image pairs, and between current and reference frames are both required in our proposed method, yet the matching methods have some differences. Stereo feature matching has an additional epipolar constraint which requires left and right image locations of the same 3D point to lie on the same horizontal line after stereo rectification and undistortion. However, as some camera calibration error is inevitable, we relax this epipolar constraint to a deviation of 2 pixels. Given a left image point (u_l, v_l), we therefore define stereo matching as the process of finding its corresponding point (u_r, v_r) in the right image based on feature descriptor similarity and subject to the epipolar geometry constraint. This is done in two steps: 1) for each (u_l, v_l), find the point with minimal feature descriptor Hamming distance within the rectangle area $u_r \in (u_l - width/2, u_l + width/2), v_r \in (v_l - 2, v_l + 2)$ as initial matching result; 2) use GMS algorithm proposed by Bian et al. (2017) to detect and remove mismatches. This procedure is fast compared with other outlier detection methods such as RANSAC. To further evaluate the quality of stereo matching, a parameter $q = \lambda\delta + (1 - \lambda)d$ is defined, where δ is the normalized reciprocal of the difference of the y-axis and d is the normalized reciprocal of the descriptor Hamming distance. After many trials, we found that prioritizing the feature descriptor similarity, specifically with a value $\lambda = 0.2$, gives optimal performance in identifying stereo point pairs.

For feature matching between frames in a temporal sequence, the epipolar constraint no longer exists, so λ is set to 0. Initial matching is built by brute force matching and followed by GMS to remove mismatches. The output may retain features which do not belong to the target and would induce motion distraction. Therefore, in the next subsection we introduce the method for combining 2D target tracking and these matched features in time sequence frames.

3.3.2. Reference Frame Selection and Outliers Rejection

Unlike common SLAM applications, which focus more on building an accurate map for the surrounding scenario, the VST- AR focuses more on the accuracy of estimated camera pose in the current frame as it first arrives. Though we designed an independent thread to optimize map and camera pose of keyframes, it only helps to reduce the pose accumulative error of the subsequent images. We found it is vital to choose an appropriate reference frame. Instead of setting the previous frame automatically as reference for the current frame, as in ORB-SLAM, we set the origin keyframe, the latest keyframe, and the previous frame as three most likely reference frame candidates. We then set the target bounding box on these frames as tracking target of SiamMask and current frame as searching image. We choose the version with highest tracking score as final reference frame. This is executed in three threads simultaneously.

Once the reference frame is selected, feature points matching between current

left frame and reference frame are calculated by performing temporal feature matching. Most outliers are removed by the GMS algorithm. If the matched features are sufficient, the tracking is deemed successful, and the pose of current frame can be retrieved by a weighted ICP method introduced in the motion estimation subsection 3.3.3. If tracking fails, we initially continue to the next frame. If five consecutive frames fail to track, all keyframes are searched to relocate.

3.3.3. Motion Estimation

For a given frame, the coordinates of observed feature points in the corresponding camera coordinate system (CCS) can be solved by triangulating the stereo pair $X = (u_l, v_l, u_r, v_r)$. Here we choose the linear triangulate algorithm proposed by Hartley and Zisserman (2003), which more accurate than the naive triangulate method used in ORB-SLAM2, because the latter assumes an ideal epipolar constraint of stereo pairs. In our system, the CCS of the first input frame is selected as the WCS, so that its observations also comprise the initial global map. This global map is updated in each of the subsequent keyframe. As mentioned, we seek the transformation $\mathbf{T}$ that maps the current frame to the WCS. This is achieved by pointcloud registration between current observations(i.e. feature points observed in the current frame) and map points in the selected keyframe. The situation is illustrated in Fig. 4. Current observations are indicated with green points and the map points belonging to the keyframe are appear in blue. Red points are map points with a match identified among current observations, and are thus the points of Interests. These corresponding points are identified using feature matching based on point feature descriptors. Each such point has coordinates K_n in the keyframe coordinate system, $\mathbf{M}_n = \mathbf{T}_{WK}\mathbf{K}_n$ in the WCS, and the $\mathbf{P}_n$ in the current CCS, where $\mathbf{P}_n, \mathbf{M}_n \in \mathbb{R}^3$ and $n = 0, 1, ..., N-1$, and N is the number of matched observations. The transformation T_{WK} from the keyframe to WCS is already known. The sought transformation $\mathbf{T}$ from the current CCS to the WCS can be then easily obtained by a weighted ICP, with the normalized feature matching quality parameter q_n as weights, for which the objective function is defined as:

$$\{\mathbf{r}, \mathbf{t}\} = \underset{\mathbf{r},\mathbf{t}}{\operatorname{argmin}} \sum_{n=0}^{N-1} q_n \left\| \mathbf{M}_n - \mathbf{r}\mathbf{P}_n - \mathbf{t} \right\|^2 \tag{2}$$

where the $\mathbf{r} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$ are the rotation and translation component of the rigid motion $\mathbf{T} \in SE(3)$ that converts points in camera coordinate system to WCS. This transformation is then used to guide fusion of the preoperative CT model with the intraoperative current frame in the visualization thread.

If tracking is lost for five consecutive frames, other keyframes are searched to match with the current frame and to relocate. This relocation may not succeed if the current frame contains severe motion blur. In that case, we simply continue to the next frame. It is acceptable for a few frames to be lost, since this has no influence on the subsequent tracking, and the loss can not even be noticed by human eyes with a normal frame rate.

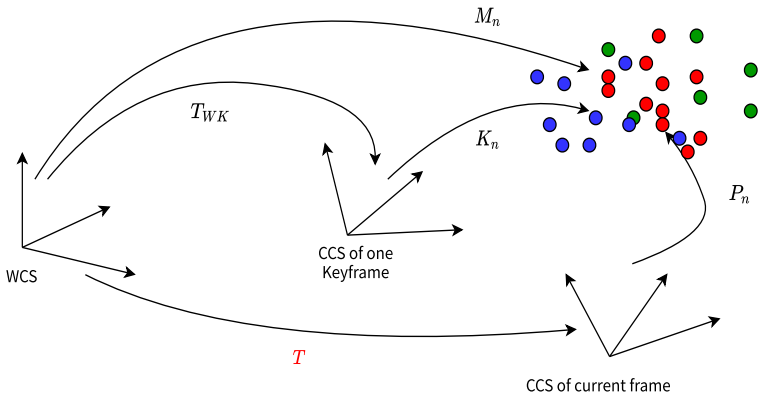

Figure 4: Illustration of coordinate systems and points used in motion estimation. Blue points are map points appearing in the selected keyframe, green points are current observations (points identified in the current frame), and red points are common observations across the two frames which have been identified using feature matching. These feature-matched points have coordinates P_n and K_n in the current and keyframes, respectively, and $M_n = T_{WK}K_n$ in the WCS. The transformation T between the current frame and WCS is computed by rigidly registering the matched points.

3.4. Mapping and Pose Optimization

As mentioned in 3.3.3, our method relies on a set of 3D points in the WCS, called the map points and their corresponding feature descriptors in the reference frame to estimate current motion of the camera. In this section we describe how to maintain the set of map points, and the database of candidate reference frames; that is, the keyframe database. Compared with outdoor application scenarios of ORB-SLAM2, the surgical scenario in our application is usually small and no scale error exists in our weighted ICP based motion estimation. We found there is therefore no need to perform loop detection for keyframes in our system, which saves memory and computation resources.

3.4.1. Map Management

The initial map is built using the first stereo frame pair. The target bounding box is manually selected in the left frame, and this frame is then set as the searching image for SiamMask. Only features inside the resulting target mask are reserved to match with features in the right image. The matched features are then triangulated and saved as map points (see Fig. 5). For each map point, the following data are recorded: 3D coordinates $\mathbf{M}_n$ in WCS; quality score q_n; number N_n of observations; IDs $\mathbf{f}_n$ of frames in which the point is observed; and the point's IDs $\mathbf{I}_n$ in each frame $\mathbf{f}_n$. The last two parameters are vectors, since the point may be detected in multiple keyframes. Once a new keyframe is inserted, tracked feature points inside its tracked mask are added into the map. N_n is updated for all points after every successful tracking. Those map points whose N_n and q_n are more than three standard deviations from the means of

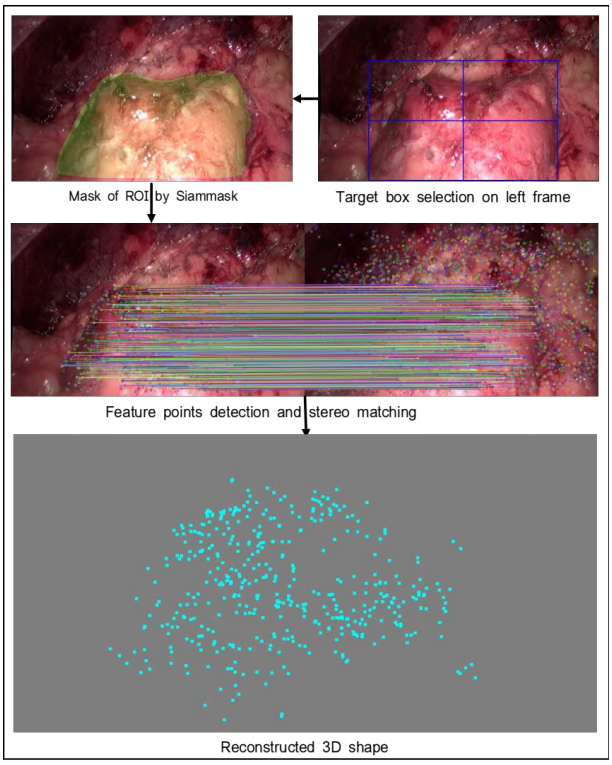

Figure 5: The reconstruction process of initial map.

all map points, or for which the related keyframes are deleted are themselves deleted to save memory.

3.4.2. Keyframes Management

Since our application scenario is relatively small in space, a small number of keyframes is enough; we set the maximum number of keyframes to 10. The first frame, in which the target bounding box is selected, is inserted as a fixed origin keyframe and never culled or updated. Any later successfully tracked frame is inserted as a keyframe only if it meets the following criteria: 1) more than 20 frames have passed from the last keyframe insertion; 2) number of tracked points is <80% of number of map points in its reference frame; and 3)number of successful stereo feature pairs generated in its target area which are then triangulated as new map points is >50. Once a frame becomes a keyframe, target tracking is executed on it by setting the target box in its reference frame as its tracking target. The output bounding box is saved, and feature points inside its output mask after outliers rejection are added to the map. The frame pose $\mathbf{T}$ with respect to WCS, the corresponding features point $\mathbf{x} = (u_l, v_l, 1)$ and feature descriptor on left image of every new map point are also saved as properties of the keyframe. To keep a fixed maximum number

of keyframes, those keyframes with <40 map points remaining, or for which $\geq 80\%$ of their map points have been observed in other keyframes are deleted. All the values of parameters mentioned here have been chosen through trial and error to ensure enough keyframes are maintained while keeping computational resources affordable.

3.4.3. Map and Pose Optimization

For each new current frame, a reference frame with respect to which pose is estimated is selected from among the set of keyframes. The accumulative error resulting from this selection should be as small as possible. BA is an efficient method to optimize both map points and frame poses by minimizing the re-projection error of all map points on their observed frames. That is, for every map point $\mathbf{M}_n$ with $N_n > 1$, the pose of the observed keyframe and related feature point are $\mathbf{T}_k$ and $\mathbf{x}_n^k$ respectively, where $k = 0, 1, ..., N_n - 1$. The BA objective function can be expressed as:

$$\{\mathbf{M}_n, \mathbf{T}_k\} = \underset{\mathbf{M}_n, \mathbf{T}_k}{\arg\min} \sum_{n,k} \rho \left(\left\| \mathbf{x}_n^k - \pi \left({\mathbf{T}_k}^{-1}, \mathbf{M}_n \right) \right\|_{\mathbf{\Sigma}}^2 \right) \tag{3}$$

where ρ is the robust Huber cost function, $\pi \left({\mathbf{T}_k}^{-1}, \mathbf{M}_n \right)$ is the function that is the function that projects $\mathbf{M}_n$ to the image plane of the camera with pose $\mathbf{T}_k$, and $\mathbf{\Sigma}$ is the covariance matrix associated with matching quality of the key point. This optimization is solved using the Levenberg–Marquardt method implemented in g2o (Kümmerle et al., 2011) after a new keyframe is added. Non-updated keyframes and map points which are added while the optimization is running are updated according to their reference frame once optimization is finished.

4. Experiments and Results

Experiments on both ex-vivo tissue phantom and clinical datasets were carried out to evaluate the performance of our proposed method. The system was implemented using C++, with the help of open-source libraries OpenCV, g2o, Pytorch and VTK. All experiments were run on a desktop computer with an Intel CPU (Intel@ Core'TM i7-7700 CPU @ 3.60 GHz) and an NVidia GTX 1060 GPU with 3 GB memory. Further results, described in relevant sections below, are provided as supplementary materials.

4.1. Experiments on Tissue Phantom Dataset

In tissue phantom experiments, a pair of porcine kidneys were used to simulate the target organ during surgery. The kidneys were scanned using a clinical CT scanner, with resulting image size $512 \times 512 \times 368$ and voxel spacing 0.6 mm $\times$ 0.6 mm $\times$ 0.5 mm, which is comparable to preoperative CT images during clinical surgery. These images were then segmented and used to build the preoperative 3D models.

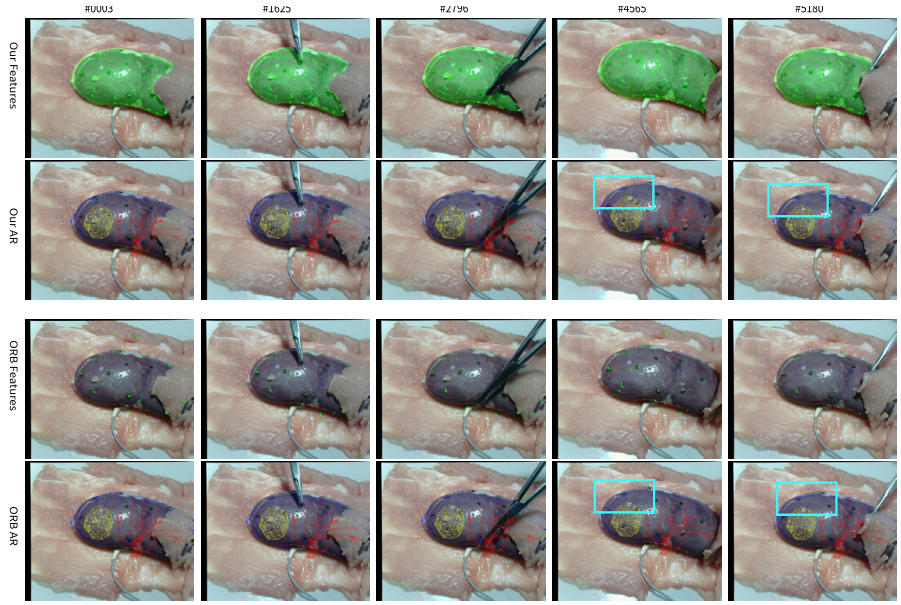

Figure 6: Comparison of the proposed method and ORB-SLAM2 on middle-level distraction videos of the tissue phantom dataset. Several representative frames are selected, with frame numbers indicated across the top. Row 1: feature points (green circles) and SiamMask segmentation (green region) identified with our method; row 2: resulting AR projections for our method, showing preoperative model points (blue), tumour surface (yellow), and vasculature (red); row 3: feature points identified with ORB-SLAM2; row 4: resulting AR projections for ORB-SLAM2 (same structures as for row 2).The obvious differences of the projections of these two methods can be seen inside the cyan boxes of raw 2 and raw 4.

Videos of the kidneys were captured by a USB binocular camera KS8A17-3.0AF produced by Shenzhen Kingsen Technology co., Ltd with a 5 cm baseline and 2560×960 frame size to imitate clinical binocular laparoscopes. To simulate the real operation environment as much as possible, 12 videos were captured. Each video contained about 8000 frames. Two of the videos recorded scenes in which only the kidney was moving, to simulate target motion in surgery due to respiratory or cardiac motion. Four videos contained motion of both the kidney and the surgical tools. Motion distractions from both surgical tools and surrounding tissues were present in the remaining videos.

To evaluate the accuracy of the motion estimation, a magnetic tracking system was used to record the motion of the kidney when capturing the videos. A sensor of the magnetic tracking system was attached to the kidney, and the organ motion so captured was considered as gold standard. Both magnetic tracking system and binocular camera remained still during filming, so that recovery of target motion only could be assessed. The transformation between camera and magnetic tracking coordinate systems was found by aligning the point clouds of the target organ obtained by each system.

Here, we compared the performance of the stereo version of ORB-SLAM2

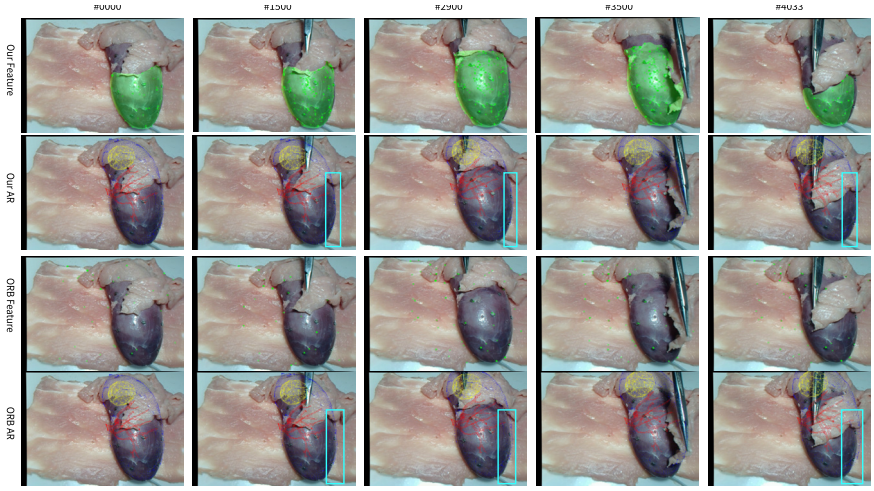

Figure 7: Comparison of the proposed method and ORB-SLAM2 on high-level distraction videos of the tissue phantom dataset. Several representative frames are selected, with frame numbers indicated across the top. Row 1: feature points (green circles) and SiamMask segmentation (green region) identified with our method; row 2: resulting AR projections for our method, showing preoperative model points (blue), tumour surface (yellow), and vasculature (red); row 3: feature points identified with ORB-SLAM2; row 4: resulting AR projections for ORB-SLAM2 (same structures as for row 2). The obvious differences of the projections of these two methods can be seen inside the cyan boxes of raw 2 and raw 4.

with that of our proposed method since there are some similarities between them. Both the parameters related to ORB feature points detection algorithm, which include the number of features, the minimal and maximal threshold of FAST corner detection, and number of image pyramid levels are set to the same. The initial registration transformation from both ORB-SLAM2 and the proposed method of the same video are also set as the same.

Since the predicted transformation $\mathbf{T}^i_{\mathbf{ex}} \in SE(3)$ and the gold standard transformation $\mathbf{T}^i_{\mathbf{gt}} \in SE(3)$ from magnetic tracking system are in different coordinate systems, the rigid transformation $\mathbf{S}$ converting $\mathbf{T}^i_{\mathbf{ex}}$ to $\mathbf{T}^i_{\mathbf{gt}}$ is solved by Horn's method Horn (1987). The Absolute Trajectory Error (ATE) $\mathbf{T}^i_{\mathbf{ex}}$ can be then calculated as:

$\mathbf{E}^i := \left(\mathbf{T}^i_{\mathbf{gt}}\right)^{-1} \mathbf{S}\mathbf{T}^i_{\mathbf{ex}}$

where $i = 0, 1, ..., C$, and C is the number of frames in video.

We compute the root mean squared error (RMSE) of translation component of $\mathbf{E}$ and rotation angles around three axes (X, Y, Z) as final metrics. The errors of both our method and ORB-SLAM2 compared with the gold standard are given in Table 1. As we can see, the proposed method outperforms ORB-SLAM2 by a large margin in the highly dynamic environment scenario, while the difference is small when no motion distraction is present. This is reasonable, because ORB-SLAM2 already performs well in scenarios without motion interference, so little improvement can be made.

Table 1: The motion estimation errors of our proposed method and of ORB-SLAM2 on our tissue phantom datasets. In this table, "OUR" indicates our proposed method and "ORB" indicates ORB-SLAM2.

Errors / Sequences	Angle X(deg)		Angle Y(deg)		Angle Z(deg)		Translation(mm)	
	OUR	ORB	OUR	ORB	OUR	ORB	OUR	ORB
Without Distraction	1.68	2.86	2.50	2.95	3.35	2.28	2.31	2.55
Middle-level Distraction	3.14	2.79	1.86	1.48	4.82	3.46	3.43	3.90
High-level Distraction	2.71	3.53	2.02	4.76	3.35	3.15	3.56	6.89

For an intuitive comparison, we show the results of the key step, that is the feature points used to estimate the motion. Several typical frames from each different level videos are selected in this paper, which contains the surgical instruments interact with the target organ, occlusions from surrounding tissues. In Fig. 6 and Fig. 7 we list the images with the features used to estimate the motion together with the tracked region of target of our method and that of features detected and used in ORB-SLAM2. Moreover, the projection of preoperative models guided under the estimated motion are also showed as a quantitative metric. Compare the projections (the blue points) inside the cyan boxes, it can be seen that the projections of our method is more close to the boundary of the kidney on the images. The readers are suggested to zoom the figures for a more clear view. We can see that the region tracking method provide a good target region estimation even at the present of occlusion from surgical instruments or surrounding tissues. All the feature in our method are detected in this target erea, so that we get a more dense feature points distribution compared with ORB-SLAM2 when we set the same number of feature to detect. And the corresponding videos are provided as supplement materials.

Fig. 8 shows the trajectories of center point of tumour from gold standard, our proposed method, and ORB-SLAM2 for an example highly dynamic scenario. The ORB-SLAM2 trajectory represents the initial motion estimate for each frame, rather than the trajectory saved after global optimization, since the former are the values applicable to real-time fusion with the CT model, during the operation. Our proposed method clearly achieves a more robust trajectory estimation and has better motion consistency.

4.2. Qualitative experiment with clinical datasets

Clinical datasets were obtained from Shanghai Renji hospital. The IRB approval was obtained from the IRB board of School of Medicine, Shanghai Jiao Tong University, and its reference number is SH9H-2019-TK323-1. A set of images of a calibration chessboard were captured just before capturing the surgical environment, and all laparoscope parameters were then fixed during the surgery. These images were used to calibrate the laparoscope by the method described in 3.1. The surgery scene in which surgeons were finding and estimating the tumor

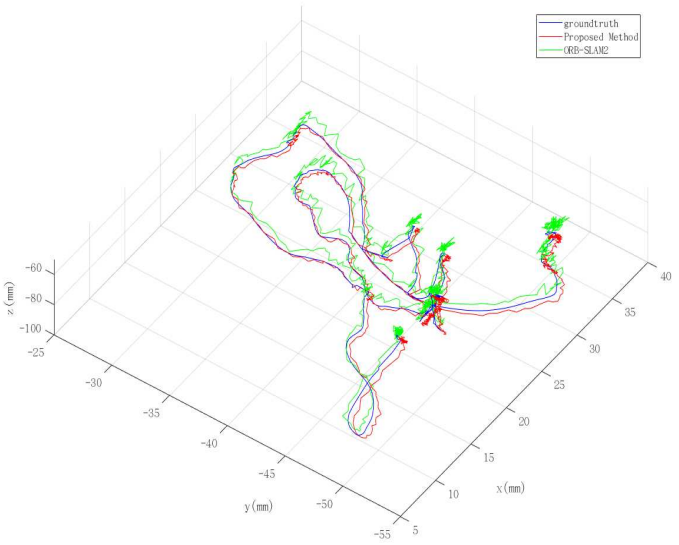

Figure 8: Trajectories of camera from gold standard, our method and ORB-SLAM2 for one high dynamic environment.

boundary was captured. These clinical datasets contain both the preoperative CT images and the corresponding intraoperative endoscope videos. The intraoperative videos contain fast camera motion, surgical instruments interaction with the target organ, and smoke generated during tumor ablation. We tested our method and ORB-SLAM2 on these clinical datasets, focusing especially on comparing their performance on those cases involving motion blur, motion distraction from surgical tools, and smoke. As there is no gold standard, only qualitative visualization results are provided.

Fig. 9 gives the performances of both methods when motion blur occurs. Here "OUR" means the proposed method and "ORB" represents ORB-SLAM2. The effective features points (that is, the projections of the visible map points in current frame) observed on current frame of both method are drawn on the original image in little green circles. Besides, the region tracking results of the proposed method are also drawn together with the feature points as light green mask. "AR" means augmented reality effects, which are actually the projection of the preoperative models guided by the estimated motion on current frame. Each row contains these four items of the same column image with its frame id on the top. While comparing items of the first row with that of the third row, we can find that the proposed method utilizes less features that don't belong to the target region while estimating the motion. The AR effect results (the second and the last row), especially the image parts inside the cyan box shows the boundary of the projections (blue points) of our method are fitted better with the kidney boundary on the images, which as a result, gives us a direct sense that the proposed method is more accurate than OBR-SLAM2. The readers are

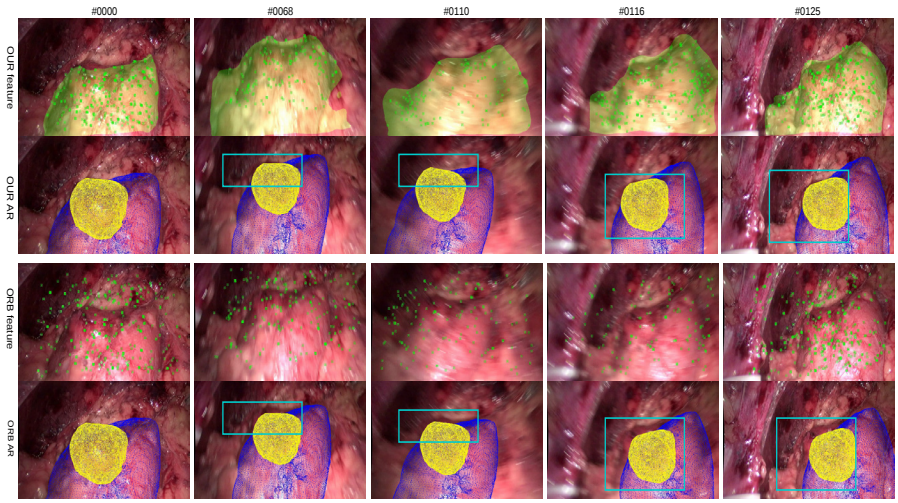

Figure 9: Comparison of the proposed method and ORB-SLAM2 on the clinical dataset in the presence of motion blur. Several representative frames are selected, with frame numbers indicated across the top. Row 1: feature points (green circles) and SiamMask segmentation (green region) identified with our method; row 2: resulting AR projections for our method, showing preoperative model points (blue) and tumour surface (yellow); row 3: feature points identified with ORB-SLAM2; row 4: resulting AR projections for ORB-SLAM2 (same structures as for row 2). The differences of the projections of these two methods can be seen inside the cyan boxes of raw 2 and raw 4.

suggested to zoom the figures for a more clear view. From the first row of Fig. 9, we can also see that the target tracking results may contain some areas which don't belong to the surface of the target organ. However, the points maintained in the map dataset are used as candidate matches for current detected features, it's a natural process to further removing the features that are not belong to the target. As a result, the proposed don't rely on a 100 percent accurate target region tracking. The worst situation is that there are too little region or even no region are tracked, in these case the motion estimation will fail. If the number of consecutive last frames reaching the threshold, the system will try to relocate.

Several frames of the performances for both methods with the presentation of motion distraction from surgical tools and smokes are shown in Fig. 10. As shown in the first row, the region tracking only givens a part of the target organ surface, however, it still estimate a relative accurate motion compared with ORB-SLAM2 when comparing the second and the last row. Comparing the AR effects of the proposed method on frame 0089, 0402 and 0461, we can find that the smoke do damage the accuracy of motion estimation, though it's more obvious on ORB-SLAM2. We found the main reason is that once the smoke occurs or becomes dense, less features can be matched with those in the map points dataset since the image patches around the feature points varied a lot. Besides, the smoke always persists longer, which would frequently trigger the relocation process or even reset the tracking system. We regard solving this

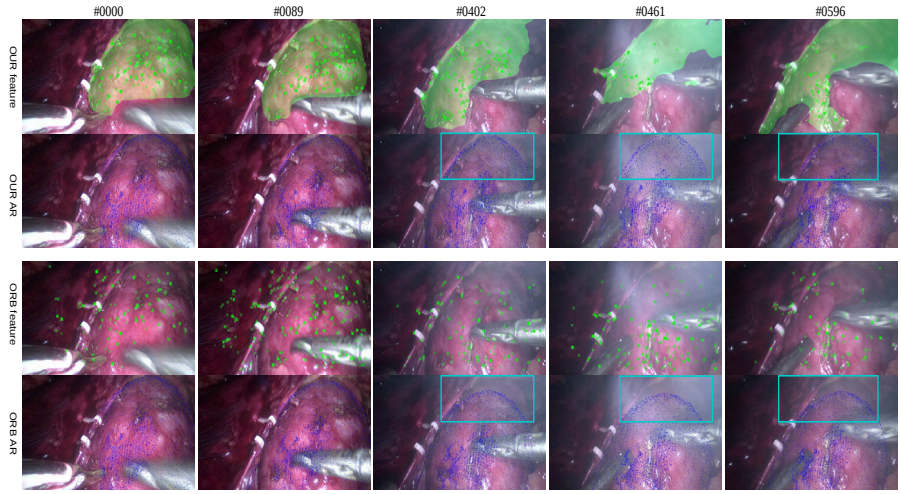

Figure 10: Comparison of the proposed method and ORB-SLAM2 on the clinical dataset with surgical tools and smoke appearing in the scene. Several representative frames are selected, with frame numbers indicated across the top. Row 1: feature points (green circles) and SiamMask segmentation (green region) identified with our method; row 2: resulting AR projections for our method, showing preoperative model points (blue); row 3: feature points identified with ORB-SLAM2; row 4: resulting AR projections for ORB-SLAM2, showing preoperative model points. The differences of the projections of these two methods can be seen inside the cyan boxes of raw 2 and raw 4.

problem as a future work.

5. Conclusion

In this paper, we proposed a real-time motion tracking method aiming to achieve robust and long term tracking in highly dynamic surgery environments. The proposed method can be regarded as a special version of ORB-SLAM that is tailored for rigid VST-AR applications in minimally invasive surgery. Our system comprises four threads running in parallel: 2D target tracking, motion tracking, map and pose optimization, and visualization. A computationally efficient target tracking and segmentation method SiamMask is used to keep focus on the target region and to exclude motion distractions from surroundings. The reference frame selection strategy is redesigned compared with ORB-SLAM to achieve more accurate motion estimation for the first calculation of each arriving frame. As our application environment is much smaller than the outdoor scenarios for which ORB-SLAM is designed, we found it is acceptable to keep only a small number of keyframes and to delete the loop closing, saving much computation. Our proposed method is tested on both ex-vivo tissue phantom datasets and clinical surgery datasets with the presence of motion blur, surgical instruments interference, and smoke. Both quantitative and qualitative results reveal an encouraging performance. However, longer persisting smoke within

the scene reduces the robustness of the proposed method, which we will tackle in future research.

6. Authors' contribution

Tingting Jia: Conceptualization, Methodology, Software, Validation, Writing-Original Draft. Zeike A. Taylor: Writing-Review & Editing, Data Curation, Visualization; Xiaojun Chen: Conceptualization, Supervision, Resources, Project Administration, Funding acquisition.

7. Acknowledgements

This work was supported by grants from National Key R&D Program of China (2017YFB1302900), National Natural Science Foundation of China (81971709; 81828003; M0019; 82011530141), the Foundation of Science and Technology Commission of Shanghai Municipality (19510712200; 20490740700), Shanghai Jiao Tong University Foundation on Medical and Technological Joint Science Research (ZH2018ZDA15; YG2019ZDA06; ZH2018QNA23), and 2020 Key Research project of Xiamen Municipal Government (No. 3502Z20201030)
Conflict of Interests: None declared.

References

Allan M, Chang PL, Ourselin S, Hawkes DJ, Sridhar A, Kelly J, Stoyanov D. Image Based Surgical Instrument Pose Estimation with Multi-class Labelling and Optical Flow. In: Navab N, Hornegger J, Wells WM, Frangi A, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. p. 331–8. URL: http://link.springer.com/10.1007/978-3-319-24553-9{_}41. doi:10.1007/978-3-319-24553-9_41.

Allan M, Ourselin S, Hawkes DJ, Kelly JD, Stoyanov D. 3-D Pose Estimation of Articulated Instruments in Robotic Minimally Invasive Surgery. IEEE Transactions on Medical Imaging 2018;37(5):1204–13. URL: https://ieeexplore.ieee.org/document/8295119/. doi:10.1109/TMI.2018.2794439.

Allan M, Thompson S, Clarkson MJ, Ourselin S, Hawkes DJ, Kelly J, Stoyanov D. 2D-3D Pose Tracking of Rigid Instruments in Minimally Invasive Surgery. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). volume 8498 LNCS; 2014. p. 1–10. URL: http://link.springer.com/10.1007/978-3-319-07521-1{_}1. doi:10.1007/978-3-319-07521-1_1.

Bernhardt S, Nicolau SA, Soler L, Doignon C. The status of augmented reality in laparoscopic surgery as of 2016. Medical Image Analysis 2017;37:66–90. URL: http://dx.doi.org/10.1016/j.media.2017.01.007. doi:10.1016/j.media.2017.01.007.

Bescos B, Facil JM, Civera J, Neira J. DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes. IEEE Robotics and Automation Letters 2018;3(4):4076–83. doi:10.1109/LRA.2018.2860039.

Besl PJ, McKay ND. Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. International Society for Optics and Photonics; volume 1611; 1992. p. 586–606.

Bouguet JY. Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm. Intel Corporation 2001;5:1–10. URL: http://www.ncbi.nlm.nih.gov/pubmed/16140533.

Chen L, Day TW, Tang W, John NW. Recent developments and future challenges in medical mixed reality. Proceedings of the 2017 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2017 2017;:123–35doi:10.1109/ISMAR.2017.29. arXiv:1708.01225.

Chen L, Tang W, John NW, Ruan T, Jun J, Wan TR, Zhang JJ. SLAM-based dense surface reconstruction in monocular Minimally Invasive Surgery and its application to Augmented Reality. Computer Methods and Programs in Biomedicine 2018;158:135–46. URL: https://doi.org/10.1016/j.cmpb.2018.02.006. doi:10.1016/j.cmpb.2018.02.006.

Feuerstein M, Mussack T, Heining SM, Navab N. Intraoperative laparoscope augmentation for port placement and resection planning in minimally invasive liver resection. IEEE Transactions on Medical Imaging 2008;27(3):355–69. doi:10.1109/TMI.2007.907327.

Hartley R, Zisserman A. Multiple view geometry in computer vision. Cambridge university press, 2003.

He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 2961–9.

Horn BKP. Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America A 1987;4(4):629. doi:10.1364/josaa.4.000629.

Kim JhH, Bartoli A, Collins T, Hartley R. Tracking by detection for interactive image augmentation in laparoscopy. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2012;7359 LNCS:246–55. doi:10.1007/978-3-642-31340-0_26.

Kong SH, Haouchine N, Soares R, Klymchenko A, Andreiuk B, Marques B, Shabat G, Piechaud T, Diana M, Cotin S, Marescaux J. Robust augmented reality registration method for localization of solid organs' tumors using CT-derived virtual biomechanical model and fluorescent fiducials. Surgical Endoscopy 2017;31(7):2863–71. doi:10.1007/s00464-016-5297-8.

Kümmerle R, Grisetti G, Strasdat H, Konolige K, Burgard W. g 2 o: A general framework for graph optimization. In: 2011 IEEE International Conference on Robotics and Automation. IEEE; 2011. p. 3607–13.

Lin B, Sun Y, Qian X, Goldgof D, Gitlin R, You Y. 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 2016;12(2):158–78. URL: http://doi.wiley.com/10.1002/rcs.1661. doi:10.1002/rcs.1661. arXiv:1504.07874.

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In: European conference on computer vision. Springer; 2014. p. 740–55.

Mahmoud N, Collins T, Hostettler A, Soler L, Doignon C, Montiel JMM. Live tracking and dense reconstruction for handheld monocular endoscopy. IEEE Transactions on Medical Imaging 2019;38(1):79–89. doi:10.1109/TMI.2018.2856109.

Mahmoud N, Grasa ÓG, Nicolau SA, Doignon C, Soler L, Marescaux J, Montiel JM. On-patient see-through augmented reality based on visual SLAM. International Journal of Computer Assisted Radiology and Surgery 2017a;12(1):1–11. doi:10.1007/s11548-016-1444-x.

Mahmoud N, Grasa ÓG, Nicolau SA, Doignon C, Soler L, Marescaux J, Montiel JMM. On-patient see-through augmented reality based on visual SLAM. International Journal of Computer Assisted Radiology and Surgery 2017b;12(1):1–11. doi:10.1007/s11548-016-1444-x.

Mur-Artal R, Montiel JMM, Tardos JD. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics 2015;31(5):1147–63. URL: https://ieeexplore.ieee.org/document/7219438/. doi:10.1109/TRO.2015.2463671.

Mur-Artal R, Tardos JD. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. IEEE Transactions on Robotics 2017;33(5):1255–62. doi:10.1109/TRO.2017.2705103.

Nicolau S, Soler L, Mutter D, Marescaux J. Augmented reality in laparoscopic surgical oncology. Surgical Oncology 2011;20(3):189–201. URL: http://dx.doi.org/10.1016/j.suronc.2011.07.002. doi:10.1016/j.suronc.2011.07.002.

Plantefève R, Peterlik I, Haouchine N, Cotin S. Patient-Specific Biomechanical Modeling for Guidance During Minimally-Invasive Hepatic Surgery. Annals of Biomedical Engineering 2016;44(1):139–53. doi:10.1007/s10439-015-1419-z.

Prisacariu V. Pwp3d: Real-time segmentation and tracking of 3d objects. International Journal of Computer Vision 2011;(2004):1–23. URL: http://www.springerlink.com/index/U1J6027609T741H5.pdf.

Puerto-Souza GA, Cadeddu JA, Mariottini GL. Toward long-term and accurate augmented-reality for monocular endoscopic videos. IEEE Transactions on Biomedical Engineering 2014;61(10):2609–20. doi:10.1109/TBME.2014.2323999.

Qiu L, Ren H. Endoscope navigation and 3d reconstruction of oral cavity by visual SLAM with mitigated data scarcity. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2018;2018-June:2278–85. doi:10.1109/CVPRW.2018.00295.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. Imagenet large scale visual recognition challenge. International journal of computer vision 2015;115(3):211–52.

Snaauw SG. Camera calibration for oblique viewing laparoscopes. Ph.D. thesis; University of Twente; 2017. URL: http://essay.utwente.nl/73351/1/Snaauw{_}MA{_}TNW.pdf.

Song J, Wang J, Zhao L, Huang S, Dissanayake G. Dynamic Reconstruction of Deformable Soft-Tissue with Stereo Scope in Minimal Invasive Surgery. IEEE Robotics and Automation Letters 2018a;3(1):155–62. doi:10.1109/LRA.2017.2735487.

Song J, Wang J, Zhao L, Huang S, Dissanayake G. MIS-SLAM: Real-Time Large-Scale Dense Deformable SLAM System in Minimal Invasive Surgery Based on Heterogeneous Computing. IEEE Robotics and Automation Letters 2018b;3(4):4068–75. doi:10.1109/LRA.2018.2856519. arXiv:arXiv:1803.02009v2.

Wang Q, Zhang L, Bertinetto L, Hu W, Torr PH. Fast online object tracking and segmentation: A unifying approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2019. p. 1328–38.

Wang R, Zhang M, Meng X, Geng Z, Wang FY. 3-D Tracking for Augmented Reality Using Combined Region and Dense Cues in Endoscopic Surgery. IEEE Journal of Biomedical and Health Informatics 2018;22(5):1540–51. doi:10.1109/JBHI.2017.2770214.

Xu N, Yang L, Fan Y, Yang J, Yue D, Liang Y, Price B, Cohen S, Huang T. Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. p. 585–601.

Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006;31(3):1116–28.

Zhang Z. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000;22(11):1330–4. URL: https://www.microsoft.com/en-us/research/publication/a-flexible-new-technique-for-camera-calibration/http://ieeexplore.ieee.org/document/888718/. doi:10.1109/34.888718.