Live Tracking and Dense Reconstruction for Handheld Monocular Endoscopy

Nader Mahmoud Elshahat Elsayed Ali, Toby Collins, Alexandre Hostettler, Luc Soler, Christophe Doignon, Jose Maria Martinez Montiel

Live Tracking and Dense Reconstruction for Hand-held Monocular Endoscopy

Nader Mahmoud[1,2], Toby Collins[1], Alexandre Hostettler[1], Luc Soler[3], Christophe Doignon[2], J.M.M. Montiel[4]

***Abstract*—Contemporary endoscopic Simultaneous Localization And Mapping (SLAM) methods accurately compute endoscope poses, however, they only provide a sparse 3D reconstruction that poorly describes the surgical scene. We propose a novel dense SLAM method whose qualities are: 1) Monocular, requiring only RGB images of a hand-held monocular endoscope. 2) Fast, providing endoscope positional tracking and 3D scene reconstruction, running in parallel threads. 3) Dense, yielding an accurate dense reconstruction. 4) Robust, to the severe illumination changes, poor texture and small deformations that are typical in endoscopy. 5) Self-contained, without needing any fiducials nor external tracking devices, therefore it can be smoothly integrated into the surgical workflow. It works as follows. Firstly, accurate cluster frame poses are estimated using the sparse SLAM feature matches. The system segments clusters of video frames according to a parallax criteria. Next, dense matches between cluster frames are computed in parallel by a variational approach that combines Zero Mean Normalized Cross Correlation (ZNCC) and a gradient Huber norm regularizer. This combination copes with challenging lighting and textures at an affordable time budget on a modern GPU. It can outperform pure stereo reconstructions because the frames cluster can provide larger parallax from the endoscope's motion. We provide an extensive experimental validation on real sequences of the porcine abdominal cavity, both in-vivo and ex-vivo. We also show a qualitative evaluation on human liver. Additionally, we show a comparison with other dense SLAM methods showing the performance gain in terms of accuracy, density and computation time.**

***Index Terms*—Endoscopy, Laparoscopy, Augmented Reality, Dense Reconstruction, Tracking, SLAM.**

I. INTRODUCTION

Minimally invasive surgical (MIS) intervention has gained substantial popularity over the past decade. Surgeons perform such interventions by manipulating an endoscope and surgical tools whose motions are controlled either by the surgeon, an assistant or a surgical robot. Recovering dense 3D information from intra-operative endoscopic images together with relative endoscope position are fundamental blocks for accurate computer-assisted guidance in MIS.

External rigid laparoscope tracking devices can provide accurate relative camera pose with respect to the Operating Room (OR), however they have limitations. The need for "line-of-sight" visibility of the optical markers requires careful planning of the tracking devices. Secondly, it requires more equipment in the OR, adds to cost, and can add to setup time because a hand-eye calibration is required. Thirdly, it cannot provide directly the camera pose relative to the internal surgical environment, which in most applications is needed. Furthermore, the optical markers are not at the tip of the scope, so pose uncertainty propagates significantly to the endpscope's tip. On the other hand, active reconstruction techniques such as structured light [1], shape-from-polarization [2], and time-of-flight [3] can recover depth and/or surface normal information without external tracking, but they require adapted endoscopic hardware, and hence have had limited practical use.

Vision-based techniques such as SLAM, has received particular attention because they can reconstruct the internal surgical environment while keeping track of the camera with respect to the internal environment, from the sole input of monocular video. Current robust monocular endoscopic SLAM approaches [4], [5] use sparse features points to recover 3D scene geometry, thus the resulting scene representation is sparsely furnished and incomplete. Recently, dense SLAM approaches [6], [7] can achieve high quality dense reconstructions in real-time for non-medical applications, however these techniques have been of limited use in endoscopy. They require constant illumination and unchanged pixel brightness with respect to the view direction. They have been experimentally proven to perform robustly for indoor scenes, however the assumptions are violated in endoscopy where the light source is attached to the endoscope tip, which produces significant illumination variability, in addition to specular reflection. In general, surgical scenes are challenging for vision based reconstruction techniques because of poor textures, occlusions, specular reflection, discontinuities and organ deformation.

In this paper, we propose a novel real-time dense SLAM system that is able to cope with the above challenges, and has been successfully applied in laparoscopy. The proposed system extends the state-of-the-art sparse ORB-SLAM [8] with a novel dense multi-view stereo-like approach. The proposed system exploits the parallel tracking and sparse reconstruction obtained by fine tuning ORB-SLAM as proposed in [9] and adds a new thread performing dense reconstruction to ORB-SLAM pipeline. It does this without interrupting the sparse SLAM threads, to maintain real time tracking. It effectively selects keyframe images, and a cluster for each keyframe of neighbor images, and computes their relative poses and a high quality dense reconstruction. The crux of the dense reconstruction is a variational approach where the data term is based on illumination invariant ZNCC instead of sum of squared differences (SSD) or sum of absolute difference (SAD) used by previous dense SLAM methods [6], [7]. Furthermore, the proposed system provides a live global and consistent dense reconstruction of the surgical scene by merging and aligning the depth maps on-line.

The obtained dense reconstruction together with the esti-

[1]IRCAD (Institut de Recherche contre les Cancers de l'Appareil Digestif), France. eng.nader.mahmoud@gmail.com

[2]ICube (UMR 7357 CNRS), Université de Strasbourg, France.

[3] IHU, Institut Hospitalo-Universitaire, Strasbourg, France.

[4] Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain. josemari@unizar.es

mated relative endoscope pose has several important applications. It enables Augmented Reality (AR) applications, e.g. [10], [11], where pre-operative/intra-operative 3D models are registered with dense reconstruction and hence allow internal organ structures to be visualized in real-time. It can also be used to compensate for breathing motion and track tissue for laser ablation. Furthermore, dense 3D reconstruction facilitates the extraction of 3D features for recognition and classification applications in gastro-endoscopy, e.g. polyp classification [12].

The remainder of this paper is organized as follows: Sec. II provides a review of the related work. Sec. III gives a detailed description of the proposed system. Sec. IV provides an experimental evaluation of the proposed system. Conclusion and future directions are presented in Sec. V.

II. RELEATED WORK

Vision-based reconstruction techniques are a key enabling technology for recovering the 3D tissue geometry from the surgical site without invasive instruments [13]. Approaches such as SLAM and Structure from Motion (SfM) have been successfully applied to a variety of anatomical settings such as gastroscopy [5], skull surgery [14], hernia repair [4] and laparoscopic organ tracking [11]. Sparse reconstruction methods consider salient image features for 3D reconstruction, and estimate camera poses by minimizing re-projection errors. The use of discriminative and rotation-invariant features [15], [16] achieves robustness to illumination changes, rotations and small deformation, that are typical in endoscopy. However, these methods poorly describe the surgical scene, because they only reconstruct features, and not dense surfaces.

A seminal work in providing SLAM-based dense surgical scene reconstruction was proposed by Mountney et al. [17]. Sparse Extended Kalman Filter (EKF) SLAM reconstructions are meshed and textured with images from a stereo-laparoscope sequence. Due to the sparse representation of the scene, artefacts are unavoidable in the final reconstruction. Totz et al. [18] have expanded stereo EKF SLAM with additional virtual features, then applied dense stereo algorithm for better describing tissue surface. EKF SLAM approaches suffer from poor scaling thus the dense reconstructions were limited to smaller regions.

In contrast, dense Multi-View Stereo (MVS) approaches attempt to recover the depth of every pixel in the images using known camera poses. Dense MVS takes a possibly very large set of images [19], [20] and can reconstruct highly detailed 3D geometry that explains the images under some assumptions e.g. rigid Lambertian surfaces, photo-consistent, known object silhouettes or shape priors [21], [22]. However, the computational complexity of estimating dense geometry with MVS has been a practical barrier to its use for real-time applications, such as computer-assisted endoscopy. There is growing interest to import MVS methods to real time constraint [23]. Recently, the unique combination of stereo-based reconstruction and Shape from Shading (SfS) in a single optimization scheme [24] allows to obtain reconstructions with varying albedo and illumination.

Newcombe et. at. [25] made a significant performance boost towards dense real-time SLAM and showed the advantage of reconstruction from large number of video frames taken from very close viewpoints, where photometric-consistency is possible. They formulated and solved an energy optimization with a photometric data term and a regularizer term to obtain a dense model. The model is further exploited to improve the tracking performance [6]. Generally, the photometric constancy assumed by [25], [6] does not hold for images captured over a wide baseline or when lighting changes significantly, which is the normal situation in endoscopy.

Recent research has been done to improve the variational approach of Newcombe et. at. [6] to handle the challenges in endoscopic scenes [26], [27]. Marecinczak et al. [26] considered the spherical color model as an illumination-invariant image representation, with a data term that relies on measuring pixels similarities. This data term is very sensitive to minor transformation, both in geometry (shifts and rotation) and in imaging conditions (noise and blurring). Chang et al. [27] considered the use of the ZNCC, proposed by [28], to gain more tolerance to camera gain or bias and provide better fidelity in textureless regions. In [27] high reconstruction accuracy was obtained to within few millimeters with real-time performance using a GPU-based implementation, but is applicable only to stereo-endoscopes. Both [26], [27] provides only local reconstruction of the visible region in either a stereo pair or a reference monocular image, but not a global and complete reconstruction of all captured regions in the scene.

Recently, Turan et. al. [29] have proposed a dense monocular SLAM method for global reconstruction of the surgical scene that fusing several depth maps obtained by SfS techniques. SfS techniques exploit the relationship between geometry, pixels intensities and scene illumination, and can recover dense 3D geometry from a single image. Although SfS has a superior performance in texture-less regions, it cannot handle surface discontinuities or spatially varying albedo which is common in endoscope images [30]. Additionally, the dense SLAM of [29] is only validated on synthetic datasets.

The closest approach to us is our previous work [31], where ORB-SLAM is used to explore the abdominal cavity and acquire a set of registered keyframes. After the exploration phase, the acquired keyframes were then processed as a set of stereo pairs, with a dense stereo algorithm run on each pair. This paper is a extension of [31] in several important ways. Firstly, we extended ORB-SLAM with a new thread performing the dense reconstruction that runs live and in parallel with ORB-SLAM tracking and mapping threads. This eliminates the wait for the abdominal exploration to finish before densification. Secondly, we select only the important keyframes, and around each keyframe a cluster of frames is automatically selected to the dense reconstruction, using a variational approach inspired from [25].

III. PROPOSED APPROACH

A. Approach Overview

We outline our approach in Fig. 1. We note that our system is applicable to any movable endoscope with a monocular camera, but here we focus on monocular laparoscopes. We assume laparosocpe is pre-calibrated with fixed intrinsic using [32] and lens distortion has been eliminated. We give the

default values for all free parameters in Sec. IV-E. The sparse keyframes-based SLAM system (ORB-SLAM) is extended with a new thread for dense reconstruction. We define the keyframe set as the selected frames used in ORB-SLAM for its Bundle Adjustment (BA) process.

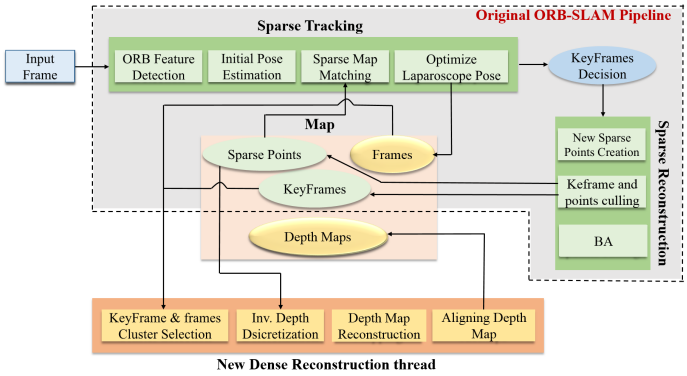


Fig. 1: System Architecture.

The dense reconstruction thread consists of four sequential modules. In the first module, we select a subset of keyframes, during live camera tracking, among all available keyframes. For each considered keyframe (I_r), a cluster of neighbor frames $\{I_1 \dots I_n\}$ is selected to have partially overlapping surface visibility (cf. Sec. III-C). In the second module, we exploit the sparse reconstruction to define the range of depths used to construct a 3D cost volume (cf. Sec. III-D4). In the third module, we perform dense reconstruction for each selected keyframe using a variational approach based on Newcombe et al. [6]. We differ by minimizing a global energy with an illumination-invariant ZNCC data term and Huber norm regularizer (cf. Sec. III-D5). In the fourth module, we obtain a globally consistent reconstruction by aligning the keyframe depth maps with the sparse SLAM map (cf. Sec. III-E). The scene is incrementally densified during the live tracking.

B. Review of Sparse Tracking and Reconstruction

1) Sparse Tracking Thread: This is responsible for frame-to-frame endoscope tracking in real-time (cf. Fig. 1). On the arrival of a new frame i, the endoscope pose $\mathbf{T}_i \in \mathrm{SE}(3)$, is roughly estimated using constant velocity motion model from the pose of last frame.

$$\mathbf{T}_i \triangleq \begin{pmatrix} \mathbf{R}_i & \mathbf{t}_i \\ 0^T & 1 \end{pmatrix} \tag{1}$$

where $\mathbf{T}_i$ is the transformation from reconstruction coordinates to camera coordinates at frame i. All map features are projected onto frame i and matched with detected ORB features. The endoscope pose $\mathbf{T}_i$ is refined by Huber $\rho_h()$ robustified non-linear minimization of the reprojection error:

$$\underset{\mathbf{T}_i}{\operatorname{argmin}} \sum_j \rho_h \left(\left\| \mathbf{x}_{i,j} - \pi(\mathbf{T}_i, \mathbf{X}_j) \right\|^2 \right) \tag{2}$$

$$\begin{aligned} \pi(\mathbf{T}_i, \mathbf{X}_j) &\triangleq h(K(\mathbf{R}_i \mathbf{X}_j + \mathbf{t}_i)) \\ h((u_j, v_j, s_j)^T) &\triangleq \frac{1}{s_j}(u_j, v_j)^T \end{aligned}$$

where $\mathbf{x}_{i,j}$ is the image point for j^{th} map point $\mathbf{X}_j$ in frame i and K is the endoscope intrinsic matrix.

2) Sparse Reconstruction Thread: This is responsible for triangulating new feature points from selected keyframes (cf. Fig. 1). The system estimates new matches across the set of keyframes and their 3D positions are triangulated. Outliers are detected and removed with strong filters, then all points and keyframes poses are further refined through BA, that minimizes eq. (2) across all keyframes. Robustness of ORB-SLAM allows it to handle the challenges in endoscopy. Points created in specular reflection and deformable regions are eliminated automatically, avoiding their negative effects on tracking performance and map corruption.

C. Frames Cluster Selection for Densification and Cluster Bundle Adjustment

Our dense reconstruction thread aims to estimate the depth of every pixel in a subset of selected keyframes. This can be computationally expensive, so we automatically choose only a subset of keyframes to densify. The selection criterion is the coverage of the current dense reconstruction in a given keyframe I_r. This reconstruction coverage is determined by projecting the current dense reconstruction to I_r, and if the reconstructed pixels fraction below 50%, I_r is selected for densification.

Upon selecting I_r, we define a cluster of n neighbor frames, $\{I_{i_1} \dots I_{i_n}\}$. The criterion for including the frames in the cluster is a measure of parallax. This is defined as the ratio between the sparse SLAM points median depth and the baseline between I_r and I_{i_n}. Frames (i.e: images and estimated poses) are stored according to their temporal location in the sequence. We then search sequentially starting from frame I_r until we find a frame whose parallax with respect to I_r exceeds a threshold α_1. This extreme frame and all intermediate frames are added to the cluster. The threshold α_1 controls the tradeoff between depth accuracy and frames overlap, where a small α_1 (i.e. small parallax) leads to noisy depths, but a higher value reduces the percentage of the overlapping pixels. It also balance the rendered parallax with photometric distortion caused by strong viewpoint change. In a second stage, frames in the cluster are reduced to remove frames from the cluster with low relative parallax, to reduce the computation cost. The condition applied is that if the parallax between frame I_{i_m} and its neighbors $I_{i_{m-1}}$ and $I_{i_{m+1}}$ is lower than a α_2 threshold, frame I_{i_m} is removed from the cluster.

The poses of the frames in the cluster are not accurate because ORB-SLAM does not perform any BA on them. Hence, we refine those poses accurately by a full BA that uses the tracked features from ORB-SLAM and minimizes eq. (3) across all the frames in the cluster and some of the other ORB-SLAM keyframes (up to 15 keyframe). The keyframes are selected as those with the most features common to I_r.

$$\underset{\mathbf{T}_i, \mathbf{X}_j}{\operatorname{argmin}} \sum_{i,j} \rho h \left(\left\| \mathbf{x}_{i,j} - \pi(\mathbf{T}_i, \mathbf{X}_j) \right\|^2 \right) \tag{3}$$

The index i ranges over all images in the frames cluster and selected SLAM keyframes, and j ranges over feature points observed by more than two cameras in the BA. The global reference is fixed during the BA to the keyframe I_r. We use Levenberg-Marquardt implemented in *g2o* [33] to carry out

that BA. The result of this computation is a set of relative poses $\{\mathbf{T}_{i_1 r} \dots \mathbf{T}_{i_n r}\}$ from I_r to $\{I_{i_1} \dots I_{i_n}\}$.

D. Reconstruction of a Keyframe's Depth Map

1) The Variational Formulation: We propose a variational energy minimization to estimate the inverse depth map $\rho(\mathbf{u}) : \Omega \to \mathbb{R}$ for a given keyframe image I_r. We use the grayscale image, denoted by $I_r : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^2$ is the 2D image domain. Our energy is the sum of a regularization term $R(\mathbf{u}, \rho(\mathbf{u}))$, and a weighted ZNCC data term $C(\mathbf{u}, \rho(\mathbf{u}))$ with the form:

$$\begin{aligned} E(\rho) &= \int_\Omega \{\lambda(\mathbf{u}) C(\mathbf{u}, \rho(\mathbf{u})) + R(\mathbf{u}, \rho(\mathbf{u}))\} d\mathbf{u} \quad (4) \\ \lambda(\mathbf{u}) &\triangleq \lambda \rho(\mathbf{u}) \end{aligned}$$

where λ is a constant and $\lambda(\mathbf{u})$ is a spatially-varying weighting factor that determines importance of the data term of pixel $\mathbf{u}$. Our empirical studies have shown that the geometrical accuracy of the recovered depth is lower for distant scene points than for closer ones because they generally have lower parallax. Thus, differently from [6], we scale the weight by $\rho(\mathbf{u})$ to reduce the data term strength for distant points.

To avoid introducing outliers in the dense reconstruction, we first detect specular reflections in I_r. This is done by thresholding saturation in HSV space with a free parameter τ. All pixels in these areas are eliminated after the optimization, because there is high uncertainty in their estimated depths.

2) ZNCC data term: In [6] a per-pixel SAD of intensity values across a cluster of images is used. In contrast, our data term is based on the ZNCC over a window around each pixel, summed for all the images in the cluster, to obtain an illumination invariant data term that can cope with the severe illumination variability in endoscopy. Each pixel $\mathbf{u} = (u, v)^T \in \Omega$ in I_r is first back-projected using $\rho(\mathbf{u})$ in the coordinate system of I_r:

$$\mathbf{X} = \pi^{-1}(\mathbf{u}, \rho(\mathbf{u})) \quad (5)$$

$$\pi^{-1}(\mathbf{u}, \rho(\mathbf{u})) \triangleq \frac{1}{\rho(\mathbf{u})} K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (6)$$

We then project $\mathbf{X}$ to each frame I_i in the cluster $\{I_{i_1} \dots I_{i_n}\}$, denoted by the 2D point $\mathbf{u}_i$:

$$\mathbf{u}_i = \pi(\mathbf{T}_{ir}, \mathbf{X}) \quad (7)$$

where $\mathbf{T}_{ir}$ is the transformation from the reference keyframe I_r to frame I_i, computed by BA in Section III-C. The data term $C(\mathbf{u}, \rho(\mathbf{u}))$ is computed by projecting pixel $\mathbf{u}$ in the reference image I_r onto $I_i \in \{I_{i_1} \dots I_{i_n}\}$ using eq. (7), and a ZNCC with correlation window size $\mathscr{W}$:

$$C(\mathbf{u}, \rho(\mathbf{u})) = \frac{-1}{n} \sum_{l=1}^{n} \mathrm{ZNCC}\left(I_r(\mathbf{u}), I_{i_l}(\mathbf{u}_{i_l})\right) \quad (8)$$

The pixels that are non-visible in all cluster frames (which project outside the image dimension) are assigned zero in the data term and eliminated after the optimization to avoid inaccurate estimation of their depths. Those ignored pixels are highly likely to be reconstructed from another reference keframe if they become visible.

3) The Regualrizer: We use a regularizer term $R(\mathbf{u}, \rho(\mathbf{u}))$. To enable a smoother reconstruction of the scene, but also to preserve depth discontinuities. This is achieved with a weighted Huber norm over the gradient of the inverse depth image:

$$R(\mathbf{u}, \rho(\mathbf{u})) = g(\mathbf{u}) \|\bigtriangledown \rho(\mathbf{u})\|_\in \quad (9)$$

where $\in$ is a free parameter of the Huber norm which determines when L^1 forming Total Variation (**TV**) or L^2 norm are used [6], to reduce the effect of the undesired stair-casing resultant from a pure **TV**. To maintain depth discontinuities across image edges, we use a per-pixel weight $g(\mathbf{u})$, with free parameter ω, to decrease the regularization strength at high gradient pixels in the reference keyframe I_r:

$$g(\mathbf{u}) = e^{-\omega \|\bigtriangledown I_r(\mathbf{u})\|_2} \quad (10)$$

4) Initialization: The ZNCC data term $C(\mathbf{u}, \rho(\mathbf{u}))$ is evaluated for keyframe I_r by means of a 3D cost volume. This has dimension M x N x ξ, where M x N is the image resolution of I_r and ξ is number of points sampling the inverse depth, that ranges between ρ_{min} and ρ_{max}. This cost volume is computed only once and an initial depth map is estimated from the cost volume by selecting $\rho(\mathbf{u})$ that minimize eq. (8) for each pixel $\mathbf{u}$. This is performed with an exhaustive search optimization over the range of inverse depths $[\rho_{min}, \rho_{max}]$.

The range $[\rho_{min}, \rho_{max}]$ is automatically defined for each keyframe from the depths provided by the sparse SLAM map. A histogram of inverse depths of all visible sparse map points in I_r (projected inside I_r) is computed, and the 20% extreme closer and farther depths are ignored, to be robust to outliers. This range is different for each keyframe depending on the depths of the visible sparse points. To include the extreme points which may have been incorrectly excluded, this interval is extended with two empirical factors β_{min} and β_{max} yielding the final interval as $[\beta_{min}\rho_{min}, \beta_{max}\rho_{max}]$. This range of inverse depths is evenly discretized into ξ sampling points.

5) Energy minimization: Eq. (4) is non-convex in the data term $\lambda(\mathbf{u}) C(\mathbf{u}, \rho(\mathbf{u}))$ and convex in regularizer term $g(\mathbf{u}) \|\bigtriangledown \rho(\mathbf{u})\|_\in$. To find a strong local minimum, we approximate the energy function with an auxiliary map $a : \Omega \to \mathbb{R}$ used to couple the two terms, as done in [6], [34]:

$$\begin{aligned} E(\rho, a) = \int_\Omega \{\lambda(\mathbf{u}) C(\mathbf{u}, a(\mathbf{u})) + \frac{1}{(2\theta)} (\rho(\mathbf{u}) - a(\mathbf{u}))^2 \\ + R(\mathbf{u}, \rho(\mathbf{u}))\} d\mathbf{u} \end{aligned} \quad (11)$$

The coupling term $\frac{1}{(2\theta)}(\rho(\mathbf{u}) - a(\mathbf{u}))^2$ enforces $\rho(\mathbf{u})$ and $a(\mathbf{u})$ to be equal as $\theta \to 0$, at which point $E(\rho, a = 0) = E(\rho)$. The global minimum of the convex term $\frac{1}{(2\theta)}(\rho(\mathbf{u}) - a(\mathbf{u}))^2 + R(\mathbf{u}, \rho(\mathbf{u}))$ is iteratively computed using primal-dual algorithm [35], [36]. At each iteration, given a solution for $\rho(\mathbf{u})$, the global minimum of the non-convex-term $\lambda(\mathbf{u}) C(\mathbf{u}, a(\mathbf{u})) + \frac{1}{(2\theta)}(\rho(\mathbf{u}) - a(\mathbf{u}))^2$ is found by performing an exhaustive search on $a(\mathbf{u})$ among the range $[\rho_{min}, \rho_{max}]$:

$$\arg\min_{a(\mathbf{u})} \lambda(\mathbf{u}) C(\mathbf{u}, a(\mathbf{u})) + \frac{1}{(2\theta)}(\rho(\mathbf{u}) - a(\mathbf{u}))^2 \quad (12)$$

The complete optimization is solved iteratively, starting at iteration $t = 1$ where θ initialized at θ^1. Both $\rho(\mathbf{u})$ and $a(\mathbf{u})$ are initialized with the initial depth map obtained from Sec. III-D4. The optimization ends when $\theta^{(t+1)} = \theta^t(1 - \varkappa t)$ exceeds a termination threshold θ_{end}. The accuracy depends on the discretization level used for the cost volume construction. To obtain a sub pixel accuracy, we perform a single Newton step proposed by [6] at each iteration.

E. Live Alignment of Keyframe Depthmaps

To obtain a globally consistent reconstruction, we combine the computed depth maps in a single coordinate frame, which is the coordinate frame of the SLAM map. Most sparse SLAM points have a corresponding 3D point in the dense maps, and we use these as anchors. The anchors are used to keep depth maps aligned with the sparse SLAM map, so that any update in the SLAM map leads to a realignment of the dense maps.

Recall that after each SLAM BA, both the sparse points and the keyframe poses are refined. This refinement may produce a misalignment of the dense maps with respect to the SLAM map. This refinement may not only involve rotation and translation but also a scale change. For this reason, we propose to align each depth map with a similarity transformation. For depth map computed from keyframe I_r, we perform a non-linear minimization of the reprojection error of sparse SLAM points P to estimate the similarity transform $\mathbf{S}$, using K neighboring keyframes that share the most feature points with I_r:

$$\underset{\mathbf{S}\in\mathbf{Sim}(3)}{\arg\min} \sum_{j\in P, i\in K} \rho_h\left(\left\|\mathbf{x}_{i,j} - \pi\left(\mathbf{T}_i, \mathbf{S}\mathbf{X}_j\right)\right\|^2\right) \tag{13}$$

where $\mathbf{x}_{i,j}$ is the image observation of sparse SLAM point j in keyframe i and $\mathbf{X}_j$ is its 3D location from the dense map of keyframe I_r, in reconstruction coordinates. Eq. (13) is repeated for every keyframe to align its corresponding depth map.

IV. EXPERIMENTAL RESULTS

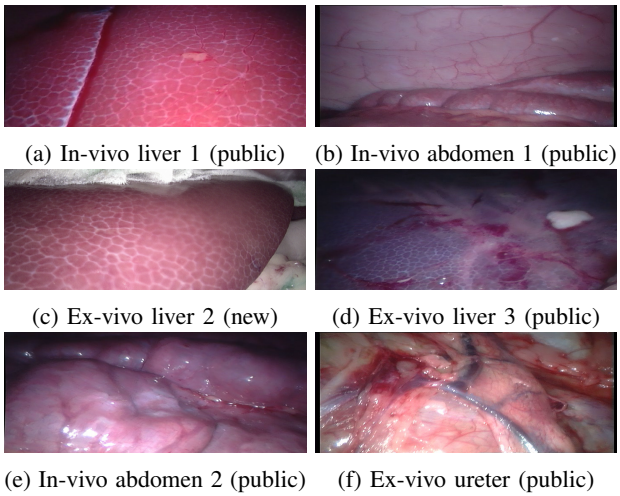
A. Benchmark Hardware and Compared Methods

The proposed system is implemented in C++ and OpenCV using a desktop computer with 8GB RAM and GeForce GTX 680 GPU with an Intel(R) Core i7 CPU 3.4GHz. We provide a quantitative evaluation of the reconstruction accuracy with respect to two ground truth methods: 1) two leading dense stereo methods [37], [27] (cf. Sec. IV-B); 2) gold-standard Computed Tomography (CT) surface (cf. Sec. IV-C). Furthermore, we compare the proposed system with the closest dense SLAM method: LSD-SLAM [7] and one of the top performing multiview stereo method where camera poses are computed from SfM [24] (cf. Sec. IV-B5). Additionally, we qualitatively evaluated the proposed system on an in-vivo exploratory sequence of human abdominal cavity (cf. Sec. IV-D). More details can be seen in the accompanying video.

B. Quantitative Evaluation Using Dense Stereo

We used the dense reconstruction of two leading stereo methods [37], [27] as our reconstruction gold-standard. According to [38] the stereo method of Chang et al. [27] is a top performing method for endoscopic images.

1) Datasets: For the evaluation, we used several sequences from Hamlyn Centre Laparoscopic/Endoscopic dataset [39] recorded by a stereo-laparoscope. Furthermore, we created a new dataset of exploratory stereo-scope camera motion at 15-20cm distance from porcine liver surface. Figure 2[a-f], shows the typical frames of the evaluation sequences. Figure 2[a,b,e] corresponds to sequences of live pigs with strong (cf. Fig. 2[a]) or small (cf. Fig. 2[b,e]) respiration. Fig. 2[c,d,f] corresponds to ex-vivo porcine sequences. The evaluation sequences had different complexities such as weak textures (cf. Fig. 2[b][e]) and repetitive textures (cf. Fig. 2[a][c][d][f]) with either smooth or strongly curved surfaces. The length of the evaluation sequences ranged between 20 seconds to 8 minutes. For each dataset, our system was used to reconstruct the scene using only images from the left laparoscope camera. To evaluate, we used their associated right images, and obtained a dense stereo reconstruction using two methods [37], [27].



(a) In-vivo liver 1 (public) (b) In-vivo abdomen 1 (public)
(c) Ex-vivo liver 2 (new) (d) Ex-vivo liver 3 (public)
(e) In-vivo abdomen 2 (public) (f) Ex-vivo ureter (public)

Fig. 2: Sample frames of the laparoscope porcine sequences used from public Hamlyn [39] and our new datasets.

2) Evaluation Metrics: Table I reports the averaged reconstruction error. Per each sequence we have varied α_1 and α_2 and yielded different reconstruction coverage, parallax, and error. The *reconstruction coverage per keyframe* is the percentage of pixels reconstructed per keyframe. A reconstructed pixel is one that is visible in all the frames of the cluster, not deleted as a specularities and not located outside the laparoscope's optical ring. *The stereo coverage* metric is the percentage of monocular reconstructed pixels for which the stereo method provide depth estimation. For each reconstructed pixel we computed the parallax rendered by the extreme frames of the cluster, and larger parallax (i.e. larger values for α_1 and α_2) leads to better reconstruction. Table I column 5 reports the average parallax among all the reconstructed pixels in all keyframes. We also report the average parallax rendered by the stereo algorithm in Table I column 6. The Root Mean Squared Error (RMSE) metric is computed as follows. We took all pixels in all keyframes for which both our method and the stereo method computed a depth estimate and measured the distance in the estimated depths. We did this with respect to both stereo methods [37] and [27]. Our reconstruction is up to scale (as with any monocular method), thus before RMSE

computation we perform a scale-only alignment by means of a Least Squares fit, where the initial guess is computed by means of a Least Median of Squares robust estimator. This monocular scale recovery is computed only once per each sequence, and then used to scale all keyframes depth maps from the same sequence. Fig 3 displays the monocular reconstruction and the stereo ground truth after the scale alignment. The Euclidean distances between all pixels from the two reconstructions are visualized in Fig. 4. We also report the standard deviation (σ) for the RMSE with respect to [27] in Table I, in addition to average reconstruction coverage and reconstruction error of LSD-SLAM except for liver 1 sequence because it has failed due to the strong respiration.

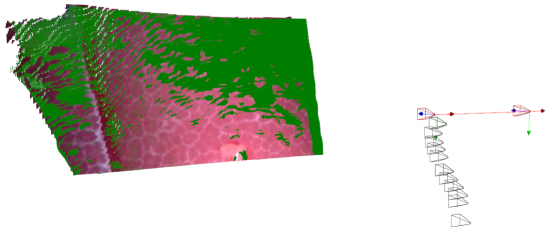
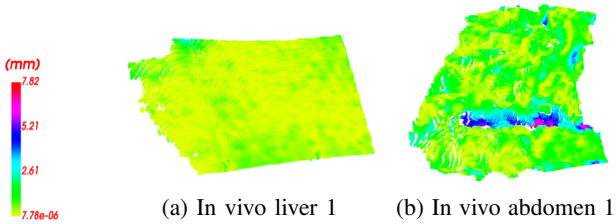


Fig. 3: Monocular (in green) to stereo (in pixel intensities) reconstruction alignment. Stereo-laparoscope cameras are shown in red with a line connecting their optical centers, and monocular frames cluster is displayed in grey.



(a) In vivo liver 1 (b) In vivo abdomen 1

Fig. 4: Euclidean distance map.

3) Results Analysis: In the same or higher monocular parallax cases with respect to the stereo methods we achieve $\leqslant 1.3 \pm 0.9$ RMSE. In such cases, it is difficult to identify whether the remaining error comes from the monocular or the stereo reconstruction. In low parallax cases, the RMSE is higher as expected. Table I also shows a superior performance of the proposed system compared to LSD-SLAM in terms of reconstruction coverage and accuracy. Fig. 5, shows the reconstruction of our system and LSD-SLAM from different points of view, where the blue frustums are the estimated camera poses at the keyframes selected by each SLAM system during the camera exploration. For in-vivo sequences the proposed method is robust to small respiration deformation as inter-frame motion in the cluster was considerably small.

4) The Influence of The Regularizer and Number of Images in The Cluster: We analyzed the effect of the regularizer in low parallax cases in Figure 6[a,d]. It shows how the correction made by the regularizer in the variational optimization is proportionally bigger in low parallax cases. It can be seen also how the RMSE is smaller in the case of the liver than in the abdomen. We conjecture that it is due to the fact that the liver surface geometry is smoother than that of the abdomen, and hence fits better the regularizer prior, which favor smooth reconstruction and because of that the final error is smaller.

TABLE I: Average reconstruction error with respect to stereo methods ([37],[27]).

Sequence	Method	Reconst. coverage per keyFrame %	Stereo coverage %	Mono pllx (*deg*)	Stereo pllx (*deg*)	Avg. RMSE (mm) [37]	Avg. RMSE (mm) [27]	σ
In-vivo liver 1	Proposed system	65	89	0.4	13.1	2.6	2.8	0.4
		66	90	5.2		1.0	1.2	0.2
		56	90	12.3		0.3	0.4	0.02
	LSD-SLAM	X	X	X		X	X	X
In-vivo abdomen 1	Proposed system	66	79	1.4	8.9	4.5	4.7	0.2
		47	88	6.1		2.9	3.3	0.2
		32	86	10.1		1.2	1.7	0.1
	LSD-SLAM	1.1	98	-		5.4	6.1	1.0
Ex-vivo liver 2	Proposed system	48	88	9.1	12	0.8	1.1	0.1
		44	79	14.9		0.7	0.9	0.1
	LSD-SLAM	1.6	85	-		2.1	2.6	1.0
Ex-vivo liver 3	Proposed system	35	98	9.8	11.4	0.6	0.7	0.1
		27	98	14.5		0.4	0.5	0.1
	LSD-SLAM	3	76	-		1.7	2.4	0.6
In-vivo abdomen 2	Proposed system	65	84	2.3	9.6	3.5	3.9	0.6
		45	95	4.8		2.9	3.3	0.6
		33	92	10.1		1.9	2.2	0.5
	LSD-SLAM	2.1	98	-		4.1	5.3	0.7
Ex-vivo ureter	Proposed system	58	82	2.9	8.5	2.0	2.3	0.6
		45	88	6		1.5	2.2	0.3
		43	90	11.7		1.0	1.9	0.2
	LSD-SLAM	1.4	92	-		2.9	3.7	1.1

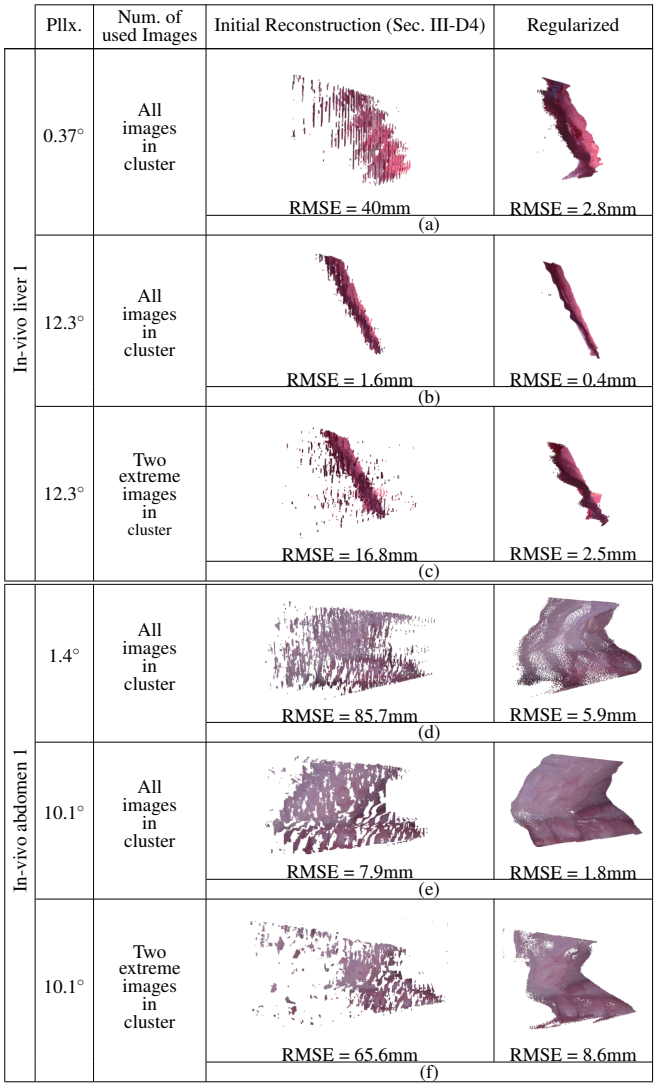
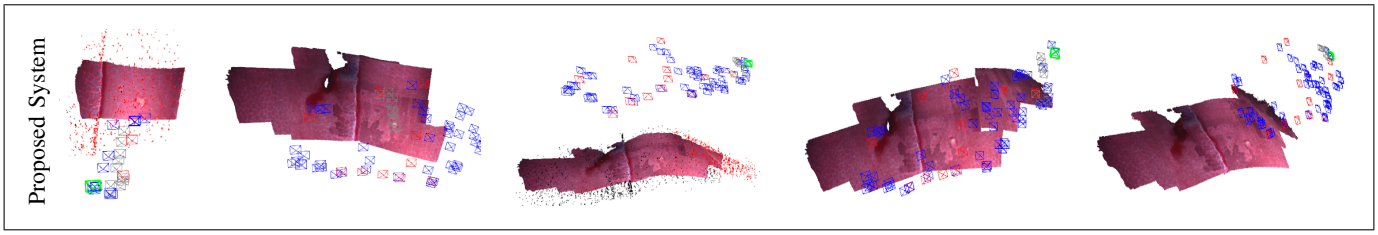
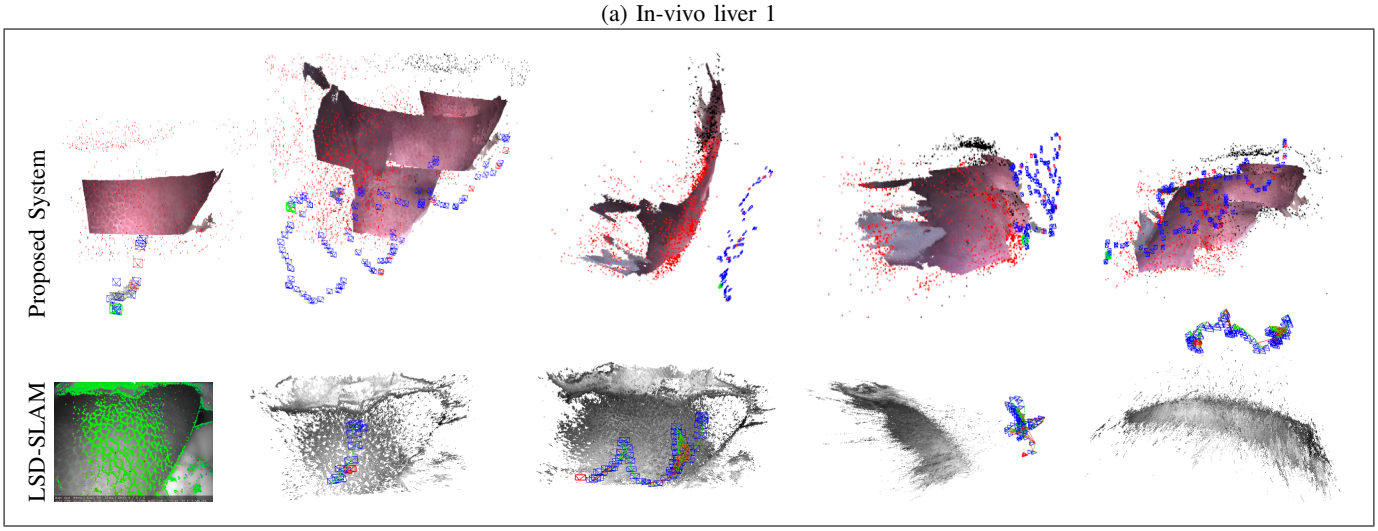


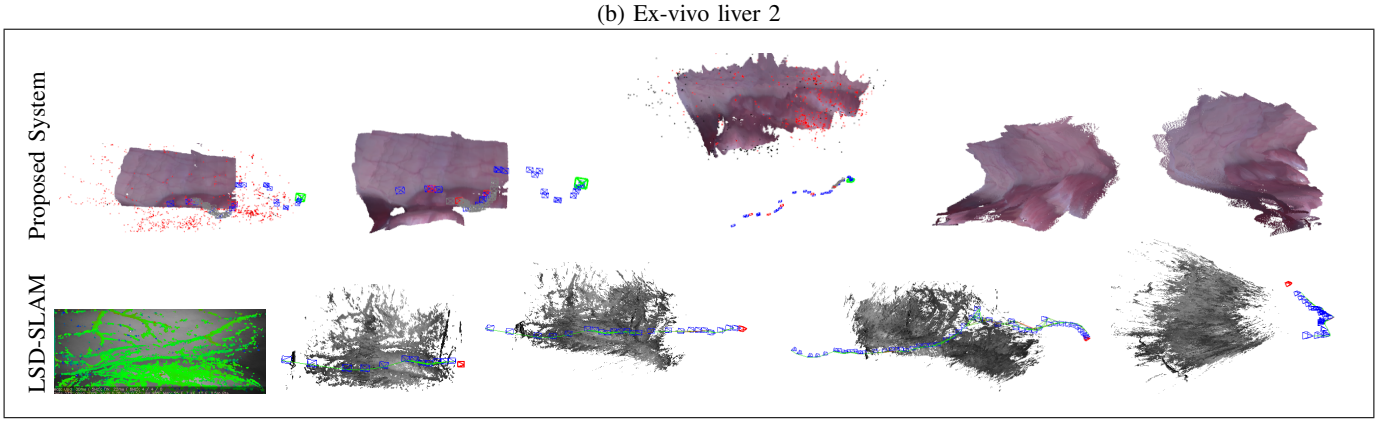
Fig. 6: Effect of the regularizer and the number of processed images in the cluster.



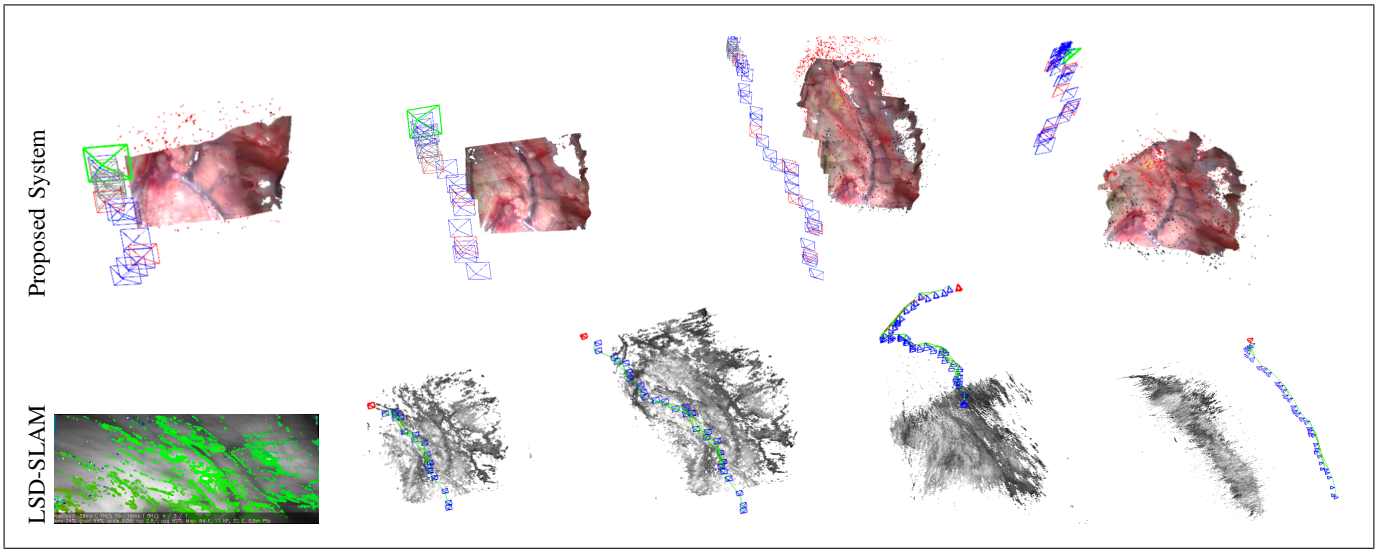
(a) In-vivo liver 1



(b) Ex-vivo liver 2



(c) In-vivo abdomen 1



(d) Ex-vivo ureter

Fig. 5: Live incremental dense reconstruction of proposed system and LSD-SLAM on different sequences visualized as points cloud. SLAM keyframes and points are colored in blue and in red, respectively. The selected keyframes used for the dense reconstruction and frames cluster are colored in red and grey, respectively. The green frustum shows the current laparoscope pose.

As expected the 3D regions with high depth gradient are reconstructed with bigger errors, it can be verified in Fig. 4(b). In high parallax cases, Fig. 6[b,e], the geometry is accurately estimated, and the regularizer effect is minimal, only to remove the stair-casing effect and provide a smoother reconstruction. The quality of the reconstruction is mostly dependent on the data term, and increasing the number of cluster images generally improve the accuracy. In Fig. 6[b,c] and [e,f] we show a comparison of the reconstruction obtained using all the frames in the cluster vs. using only the two extreme frames in the cluster. The data term is a simple two view stereo when using two images, and the lack of data constraints can lead to spurious local minimum in the variational problem. However, the cost when using a cluster of many images taken from different viewpoints generally produces a strongly constraint problem with a strong global minimum. This directly increases the chance that a good initial solution is found (cf. Sec. (III-D4)).

5) Proposed System versus Dense SfM: We evaluated the reconstruction accuracy and computation time with a state-of-the-art dense SfM method [24]. We have performed this evaluation on the ex-vivo liver 3 sequence. Figure 7[a,b] shows the final reconstruction by [24] after the filtering/refinement step. The RMSE was 0.6mm and 0.8mm w.r.t stereo methods [37], [27], respectively.

(a) (b)

Fig. 7: Dense SfM [24].

The averaged rendered parallax was 12.4 degrees. The proposed system and [24] yield similar accuracy and both render higher monocular parallax than stereo methods, however the proposed system is order of magnitude faster. The dense reconstruction of [24] took $\approx$ 4.5 min and the subsequent filtering step took $\approx$ 1.5 min.

C. Quantitative Evaluation Using CT Surface

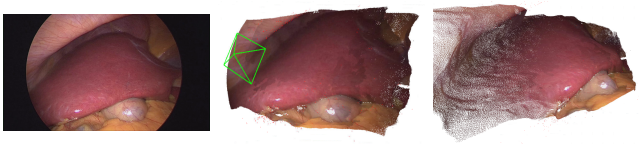
To evaluate the global reconstruction of the proposed system, we performed an ex-vivo porcine experiment with a intra-operative CT acquisition. We used a monocular laparoscope and recorded 2min ex-vivo exploratory sequence of the abdominal viscera. A sample frame of the sequence is shown in Fig. 8[a]. After the exploration is finished, a CT images was then acquired during 10 seconds expiration breath-hold and is manually segmented by an expert to generate a 3D volume with 0.876mm x 0.876mm x 0.799mm voxel size and 12,012 vertices.

We aligned the monocular reconstruction to the CT surface using a best-fitting similarity transform $\mathbf{Sim}(\mathbf{3})$. This was found by manually selecting 3 landmarks to roughly estimate $\mathbf{Sim}(\mathbf{3})$, using Horn's algorithm [40]. Then Iterative Closest Point (ICP) was run until convergence to refine the initial alignment. RMSE was then measured by the Euclidean distance of each map point to its closest point on the CT surface. Once alignment is finished, the reconstruction coverage, which is the percentage of the reconstructed CT points that were visible in the laparoscope images during exploration, is computed as follows. Firstly, we identified the visible CT surface in the estimated keyframe poses by our SLAM. The CT point is marked as visible if it is projected within at least 5 SLAM keyframe images. Then the reconstruction percentage is calculated as a ratio between number of visible CT points that are reconstructed and total number of visible CT points.

Figure 8[b,c], shows our dense reconstruction and the keyframe poses (blue frustums) estimated during the laparosope exploration. Fig. 8[e] shows the alignment between our dense reconstruction and CT model in yellow. The RMSE of our system is 1.1mm, and the width of the reconstruction is 10.4cm (cyan line in Fig. 8[e]). We show in Fig. 8[d,g] the distance error map and its cumulative distribution function map, where it can be seen that $\approx$ 84% of the points with error less than 1.5mm. Fig. 8[f] shows in green the visible CT points in at least 5 keyframes images. The coverage percentage of our reconstruction is 42% where our system only considers the overlapping pixels in the frame clusters and this percentage reduces with larger parallaxes for frame cluster selection. The RMSE of LSD-SLAM is 2.5mm with 24% reconstruction coverage. We process 80 frames of the evaluation sequence using dense SfM [24] method, that took 25min to finish and yielded the RMSE 1mm and 12% reconstruction coverage.

D. Qualitative Evaluation on In-vivo Human Liver

The visual textures of the human liver is lower than the porcine one, which makes its 3D reconstruction more challenging. To our knowledge, no dense reconstruction results have been reported on human liver with a monocular laparosocpe. We qualitatively tested the proposed system on a short sequence corresponding to an in-vivo human abdominal cavity exploration. The exploration has been done by a surgeon who has no prior knowledge about SLAM, and included fast laparoscope motion with different orientation changes. Figure 9[a] shows image sample of the patient sequence. Our sparse SLAM system was able to locate few but good sparse points to estimate laparoscope camera poses. Fig. 9[b,c] shows our dense reconstruction results from different viewpoints, which qualitatively look very promising and accurate.

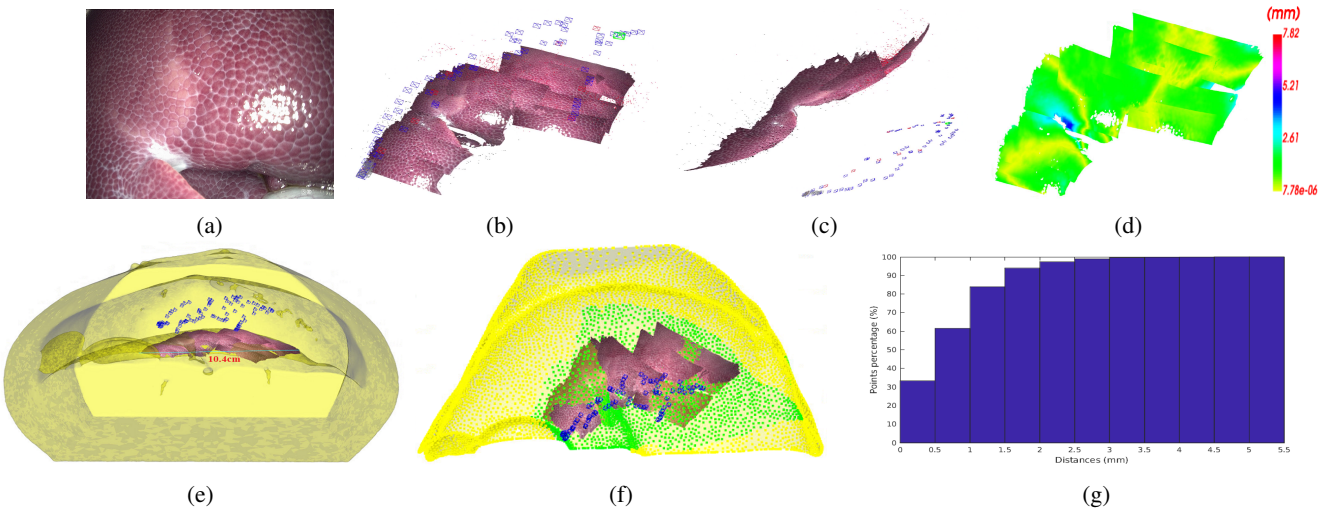


(a) (b) (c)

Fig. 9: In-vivo human liver reconstruction.

E. Free Parameters

We detail in Table II all the free parameters which were fixed during the experiments. The main sensitive parameter is the overlap between the cluster frames, controlled by α_1 and α_2. We fix $\varkappa = 0.001$ to meet a good balance between the quality and computing time trade-off in the variational minimization. Integral images were used to keep the running time invariant to the ZNCC window size as proposed in [28].



(a) (b) (c) (d) (e) (f) (g)

Fig. 8: Evaluation with respect to CT surface. (a) Sample frame. (b,c) Our reconstruction from different direction. (d) Distance error map. (e) Alignment with transparent CT volume of abdominal cavity from side view. (f) Visible CT points (green) in estimated keyframe poses (blue frustum). (g) Distance error cumulative distribution function.

TABLE II: Parameters settings.

α_1	α_2	θ^1	$\varkappa$	θ_{end}	λ	ω	$\mathscr{W}$	β_{min}	β_{max}	ξ	$\in$	τ
0.2	0.01	0.2	0.001	0.0005	0.5	0.01	19	0.8	5	51	0.001	30

F. Processing Time

We report in Table III the average execution time needed by each step of the proposed system, for dense reconstruction, and the average execution time of the two parallel threads from ORB-SLAM (Sparse Tracking and Sparse Reconstruction) for different image resolutions.

TABLE III: Average Processing Time (In Seconds).

Image Resol.	Sparse Tracking	Sparse Reconst.	Dense Reconstruction					
			Cluster selection	BA	Inverse depth Discretiz.	Cost volume	Variational minimiz.	Depth maps realignment
720x288	0.03	0.6	0.17	1.3	0.00036	3.4	6.2	0.38
960x260	0.04	0.69	0.21	2	0.0039	5.2	8.4	0.47

In the *Dense Reconstruction* thread for image resolution 720x288 of public dataset, the selection of the reference keyframe and its frames cluster took $\approx$ 0.17s followed by a Bundle Adjustment, that accurately estimates the poses of frames in the cluster $\approx$ 1.3s. It is worth noting that most of this time is spent computing the sparse matches between the frames in the cluster, the BA stage just took $\approx$ 100ms. The ZNCC cost volume construction took $\approx$ 3.4s implemented on the GPU and the cluster size varied between 5-18 frames. The equivalent time using CPU implementation varied between 18-25s. The variational solver was implemented on the CPU, yielding a computation time of $\approx$ 6.2s. Using a GPU implementation as proposed in [6] could reduce this time significantly. The depth maps re-alignment stage took $\approx$ 380ms on average. In case of our new dataset that has 960x260 image resolution, the processing time are slightly increased due to large number of images features.

V. Conclusion and Future Work

A novel real-time dense SLAM system has been presented that is able to track the endoscope at frame-rate using image features, and is able to produce in few seconds a high quality dense reconstruction of the surgical scene. The proposed system uses the sole input of monocular videos and does not need any fiducials nor external trackers, thus can be integrated smoothly into the current surgical workflow.

It has been validated and evaluated on real porcine laproscope sequences from public and our new datasets and shows a robustness to severe illumination changes and different scene textures. It also shows a very promising dense reconstruction of human liver. On one hand the accuracy of the dense reconstruction has been evaluated with respect to gold-standard CT surface and yielded 1.1mm of accuracy. On the other hand, the evaluation with respect to the stereo methods provides a similar measure of the accuracy in the laparoscope pose estimation with respect to the surface because we only apply a scale alignment, then it is mainly testing the accuracy of the camera rotation and translation with respect to the estimated surface.

Our experiments have also shown that when the camera loops back to regions already included in the map, the SLAM algorithm succeeded. The main effect of the subtle tissue deformations is that new map points are created because the old ones generate matching hypothesis that are rejected as they fail to pass the rigidity test. The net effect in the overall SLAM performance is very low. Our focus has been the fast and accurate surface geometry reconstruction, thus, we have neglected pixels with high uncertainties such as specularities and pixels not observed in all the frames in the corresponding cluster. Similarly, our algorithm uses the pixel intensities for the dense surface representation, which is a basic one unable to remove the artefacts in the reconstructed surface textures caused by the differences in illumination between the keyframes used for surface estimation. A nice venue for future

work would be to devise a better approach to interpolate the reconstruction in the areas corresponding to high uncertainty and to blend seamlessly the textures in the 3D surface as they are relevant for high quality AR visualization.

The main limitations of the proposed system are: 1) It cannot deal with very homogeneous soft-tissue surfaces that completely lacks texture characteristics. 2) It cannot deal with the strong deformations that exist during surgical manipulation. However, it has proven a robustness for small deformations caused by respiration and can robustly provide the prerequisite template for non-rigid methods such as shape from template [11], [41]. 3) It requires offline camera calibration and any change in the calibration parameters can strongly affects the system performance, and thus a technique for detecting and handling these changes online is beneficial. Future research directions may focus on a fusion with additional visual cue such as shading, that explicitly models the reflectance properties of the surface.

ACKNOWLEDGMENT

This work is part of a project of the Investissements d'Avenir program ("Investing in the Future") called 3D-Surg, funded by BPIfrance. It is also partially funded by the Spanish government DPI2015-67275-P and Aragonese DGA T04-FSE.

ETHICAL APPROVAL

Our experimental study was conducted according to a protocol received full approval from the local Ethical Committee for animal use and care (ICOMETH; protocol n 38.2015.01.069, acronym ETICA) and approved by the French Ministry of Superior Education and Research (MESR) under the reference number (2015092210412678 v4 APAFIS#1830). We have also followed the ARRIVE guidelines [42], French laws and the directives of the European Community Council (2010/63/EU).

REFERENCES

[1] J. Lin, N. T. Clancy, and D. S. Elson. An endoscopic structured light system using multispectral detection. *IJCARS*, 10(12):1941–1950, 2015.

[2] S. E. Martinez-Herrera, A. Malti, O. Morel, and A. Bartoli. Shape-from-polarization in laparoscopy. In *IEEE Symposium on Biomedical Imaging*, pages 1412–1415, 2013.

[3] T. Köhler et al. Tof meets rgb: Novel multi-sensor super-resolution for hybrid 3-d endoscopy. In *MICCAI*, pages 139–146, 2013.

[4] Ò. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. M. M. Montiel. Visual slam for handheld monocular endoscope. *IEEE Trans. on Medical Imaging*, 33(1):135–146, 2014.

[5] D. Sun, J. Liu, et al. Surface reconstruction from tracked endoscopic video using the structure from motion approach. In *Augmented Reality Environments for MICCAI*, pages 127–135, 2013.

[6] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *ICCV*, pages 2320–2327, 2011.

[7] J. Engel, T. Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, pages 834–849, 2014.

[8] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards. Orb-slam: A versatile and accurate monocular slam system. *IEEE Trans. on Robotics*, 31(5):1147–1163, 2015.

[9] N. Mahmoud, I. Cirauqui, A. Hostettler, et al. Orbslam-based endoscope tracking and 3d reconstruction. In *CARE-MICCAI*, pages 72–83, 2017.

[10] G. A. Puerto-Souza, J. A. Cadeddu, and G. L. Mariottini. Toward long-term and accurate augmented-reality for monocular endoscopic videos. *IEEE Trans. on Biomedical Engineering*, 61(10):2609–2620, 2014.

[11] T. Collins, A. Bartoli, N. Bourdel, and M. Canis. Robust, real-time, dense and deformable 3d organ tracking in laparoscopic videos. In *MICCAI*, pages 404–412, 2016.

[12] P. Mesejo, D. Pizarro, et al. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE Trans. on Medical Imaging*, 35(9):2051–2063, 2016.

[13] L. Maier-Hein, P. Mountney, A. Bartoli, et al. Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Medical Image Analysis*, 17(8):974 – 996, 2013.

[14] D. J. Mirota, H. Wang, R. H. Taylor, et al. A system for video-based navigation for endoscopic endonasal skull base surgery. *IEEE Trans. on Medical Imaging*, 31(4):963–976, 2012.

[15] David G. Lowe. Distinctive image features from scale-invariant keypoints. *ICCV*, 60(2):91–110, 2004.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, pages 2564–2571, 2011.

[17] P. Mountney and G. Z. Yang. Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. In *IEEE Engineering in Medicine and Biology Society*, pages 1184–1187, 2009.

[18] J. Totz, P. Mountney, et al. Dense surface reconstruction for enhanced navigation in mis. In *MICCAI*, pages 89–96, 2011.

[19] C. H. Esteban, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *PAMI*, 30(3):548–554, 2008.

[20] Z. Zhou, Z. Wu, and P. Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *CVPR*, pages 1482–1489, 2013.

[21] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, volume 1, pages 519–528, 2006.

[22] P. Heise, B. Jensen, et al. Variational patchmatch multiview reconstruction and refinement. In *ICCV*, pages 882–890, 2015.

[23] J. Y. Chang, H. Park, I. K. Park, K. M. Lee, and S. U. Lee. Gpu-friendly multi-view stereo reconstruction using surfel representation and graph cuts. *Computer Vision and Image Understanding*, 115(5):620–634, 2011.

[24] F. Langguth, K. Sunkavalli, S. Hadap, and M. Goesele. Shading-aware multi-view stereo. In *ECCV*, 2016.

[25] R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In *CVPR*, 2010.

[26] J. M. Marcinczak and R-R Grigat. Total variation based 3d reconstruction from monocular laparoscopic sequences. In *Abdominal Imaging. Computational and Clinical Applications*, pages 239–247, 2014.

[27] P.-L. Chang, D. Stoyanov, et al. Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. In *MICCAI*, pages 42–49, 2013.

[28] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G-Z Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *MICCAI*, pages 275–282, 2010.

[29] M. Turan, Y. Almalioglu, H. Araujo, et al. A non-rigid map fusion-based direct slam method for endoscopic capsule robots. *Journal of Intelligent Robotics and Applications*, 1(4):399–409, 2017.

[30] T. Collins and A. Bartoli. 3d reconstruction in laparoscopy with close-range photometric stereo. In *MICCAI*, pages 634–642, 2012.

[31] N. Mahmoud, A. Hostettler, T. Collins, L. Soler, C. Doignon, and J. M. M. Montiel. SLAM based quasi dense reconstruction for minimally invasive surgery scenes. *ICRA workshop C4 Surgical Robots.*, pages 36–39, 2017.

[32] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 22(11):1330–1334, 2000.

[33] R. Kmmerle, G. Grisetti, H. Strasdat, et al. G2o: A general framework for graph optimization. In *ICRA*, pages 3607–3613, 2011.

[34] A. Concha, W. Hussain, et al. Incorporating scene priors to dense monocular mapping. *Autonomous Robots*, 39(3):279–292, 2015.

[35] J.-F. Aujol. Some first-order algorithms for total variation based image restoration. *Journal of Mathematical Imaging and Vision*, 34(3):307–327, 2009.

[36] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[37] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 30(2):328–341, 2008.

[38] L. Maier-Hein, A. Groch, A. Bartoli, et al. Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. *IEEE Trans. on Medical Imaging*, 33(10):1913–1930, 2014.

[39] P. Mountney, D. Stoyanov, and G. Z. Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine*, 27(4):14–24, 2010. Available at http://hamlyn.doc.ic.ac.uk/vision/.

[40] Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987.

[41] A. Bartoli, Y. Gerard, F. Chadebecq, T. Collins, and D. Pizarro. Shape-from-template. *PAMI*, 37(10):2099–2118, 2015.

[42] C. Kilkenny, W. J. Browne, et al. Improving bioscience research reporting: The arrive guidelines for reporting animal research. *PLOS Biology*, 8(6):1–5, 2010.