

# Tracking monocular camera pose and deformation for SLAM inside the human body

Juan J. Gómez Rodríguez, J.M.M. Montiel, *Member, IEEE* and Juan D. Tardós, *Fellow, IEEE*

**Abstract**—Monocular SLAM in deformable scenes will open the way to multiple medical applications like computer-assisted navigation in endoscopy, automatic drug delivery or autonomous robotic surgery. In this paper we propose a novel method to simultaneously track the camera pose and the 3D scene deformation, without any assumption about environment topology or shape. The method uses an illumination-invariant photometric method to track image features and estimates camera motion and deformation combining reprojection error with spatial and temporal regularization of deformations. Our results in simulated colonoscopies show the method’s accuracy and robustness in complex scenes under increasing levels of deformation. Our qualitative results in human colonoscopies from Endomapper dataset show that the method is able to successfully cope with the challenges of real endoscopies: deformations, low texture and strong illumination changes. We also compare with previous tracking methods in simpler scenarios from Hamlyn dataset where we obtain competitive performance, without needing any topological assumption.

## I. INTRODUCTION

Visual Simultaneous Localization and Mapping (SLAM) and Visual Odometry (VO) in static environments have been hot research topics in the last decades and many methods have raised to solve them with outstanding accuracy and robustness using features [1], direct methods [2], or hybrid techniques [3]. The increasing popularity of these techniques has raised expectations to solve SLAM in more complex scenarios. For example, one can think of many useful applications of SLAM in Minimal Invasive Surgery (MIS) like guiding surgeons through augmented reality annotations towards the place where a polyp was detected in a previous exploration, and automatic polyp measurement to analyze its evolution. Moreover, surgical robots would greatly benefit from SLAM inside the human body as they will be more secure, robust and accurate, and they will be able to combine information coming from previous explorations or from other sensors like Computerized Tomography (CT).

However, visual SLAM inside the human body poses tremendous challenges like weak texture, changing illumination, specular reflections, and lack of rigidity (Fig. 1). Weak texture and specular reflections hinder data association algorithms based on feature matching, preventing methods like ORB-SLAM3 [4] from working in these sequences. On the other hand, changing illumination puts direct methods like DSO [2] or DSM [5] and hybrid methods like SVO [3]

This work was supported by EU-H2020 grant 863146: ENDOMAPPER, Spanish government grant PGC2018-096367-B-I00 and by Aragón government grant DGA.T45-17R and PhD scholarship of J. J. Gómez-Rodríguez.

The authors are with the Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, María de Luna 1, 50018 Zaragoza, Spain. E-mail: {jjgomez, josemari, tardos}@unizar.es.

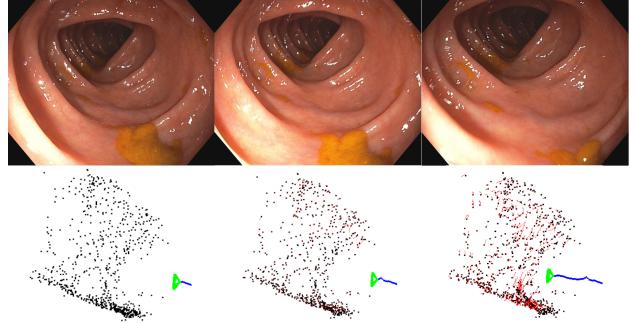


Fig. 1: Tracking the camera pose and scene deformation in a human colonoscopy from Endomapper dataset. Top: images from the sequence. Bottom: camera trajectory in blue, undeformed map points in black and map point deformation trajectories in red.

in serious trouble, as they assume constant illumination of the environment. In contrast, we solve data association with a modified Lucas-Kanade algorithm, first presented in [6], that is able to cope with local illumination changes.

But the major challenge to be addressed for SLAM inside the human body is deformable scenes, as breaking up with the rigidity assumption impairs both environment reconstruction and camera tracking. The recent DefSLAM [7] is the first monocular deformable SLAM system able to perform tracking and mapping, but it strongly relies on the assumption of a smooth continuous shape with planar topology, which does not hold in colonoscopies (see Fig. 1).

In this paper we present the first pure monocular method able to initialize a map and track camera pose and scene deformation in general scenes inside the human body (Fig. 1), without any topological or shape assumption. Our main contribution is a simple formulation that combines photometric feature tracking and an optimization based on reprojection error with spatial and temporal regularizers that encode local assumptions over the environment deformation, endowing our algorithm with enough expressivity to model complex scenes and track their deformations in real-life endoscopies. We provide quantitative evaluation on realistic colonoscopy simulations [8] and qualitative results on real human colonoscopies from the Endomapper project [9], that were out of reach for previous techniques. We present quantitative comparisons in almost-planar scenes from Hamlyn dataset [10] where we obtain competitive performance, despite not using any assumption on the scene topology or shape.

## II. RELATED WORK

The computer vision and robotics communities have developed excellent rigid visual SLAM systems [2] [3] [4] in the last years. While all these algorithms use quite different techniques they all rely in a vital, yet simple, assumption: that the environment is static. In contrast, deformable SLAM, which completely breaks up with the rigidity assumption, is still a challenging research topic.

Many works have tried to solve deformable SLAM by using sensors that provide complete 3D information of the environment like stereo or RGB-D cameras. This is the case of the seminal DynamicFusion [11] which uses RGB-D images to reconstruct highly deforming environments with an Iterative Closest Point (ICP) algorithm and a spatial regularizer to constrain deformations of points that are close to each other, that we adopt in our work. Several extensions to DynamicFusion have been developed since then, being the most notably VolumeDeform [12] which combines the use of SIFT features and reprojection error with a dense ICP data term, to reduce drift and increase robustness. In [13] they formulate a variational method that takes RGB-D images to reconstruct a deforming environment. This work is later extended in [14] introducing camera pose computation.

There is increasing interest in SLAM in Minimally Invasive Surgery (MIS), where RGB-D sensors are not available. For this reason, works like [15] [16] use depth coming from stereo images and an error function that combines reprojection errors and regularizers to perform deformable SLAM. As before, the reprojection error is augmented using other 3D terms like ICP errors or Point-to-Plane errors. Regarding the regularizers used, they are similar to the one introduced in DynamicFusion to represent that deformations occur locally, using pair-wise deformation terms between close points, which prevents divergence of individual points in the reconstruction.

Nevertheless stereo cameras are not appropriate for certain applications like colonoscopies where two cameras with enough baseline may not fit in the body cavities. In this kind of scenarios, only monocular deformable SLAM can be performed. This is even a harder problem since no real 3D information is available from a single view, scale is unobservable, and combining multiple views of a deforming scene is an open issue. The first monocular deformable SLAM system is DefSLAM [7] which splits the deformable SLAM problem in 2 threads, one for tracking and one for mapping. They use ORB features and minimize a reprojection error term with a deformation energy term that penalizes stretching and bending of the imaged surface. However, as ORB features are quite unstable in intracorporeal sequences, SD-DefSLAM [6] extends DefSLAM to a semi-direct method by integrating an illumination-invariant Lucas-Kanade tracker to perform data association, achieving better robustness and accuracy. Crucially, both methods assume that the surface has planar topology, and model the surface with a triangle mesh which impose a strong global condition over the environment: the imaged surfaces have to be continuous with no holes. This

is quite a strong assumption that seriously limits the kind of scenes that can be handled by both algorithms excluding, for example, colonoscopies (Fig. 1).

To tackle this limitation, [17] proposes a fully photometric algorithm to track camera pose and deformation using sparse 3D surfels (surface elements) under the assumption of local isometry. The use of surfels that have no constraints between them allows to model any kind of topologies. While obtaining very good results in medical scenes, the method still requires 3D information coming from a stereo camera to initialize the surfels. Also, the use of large surfels (in practice, square patches of  $23 \times 23$  pixels in the image) can easily violate the local isometry assumption and is inefficient as using too many close pixels provide redundant information with little to no improvement in accuracy [2]. Furthermore, the regularizers used impose small deformations with respect to a pose at rest, which can be inappropriate in many applications.

In contrast, in this work we propose a pure monocular method for tracking camera pose and deformation. Under the assumption of slow deformations, we perform fully automatic monocular map initialization to obtain a first seed of the environment structure. Following previous works [3], [6], we use photometric feature tracking for robustness and accuracy, and reprojection error for convergence and efficiency during optimization. In addition, we integrate two regularizers that encode our assumptions of smooth and slow deformation in order to constrain the reconstruction problem.

## III. DEFORMABLE TRACKING

This section is devoted to presenting our tracking algorithm. We first present the assumptions that govern our system. Afterwards we introduce our data association for tracking. Then we explore our algorithm for monocular map initialization and finally we present our optimization backbone and its formulation encoding each one of our assumptions.

### A. Assumptions

The biggest difficulty when dealing with deformable scenarios is that the rigid assumption is violated. This makes camera pose and deformation prediction a non-separable problem for which infinite solutions arise, i.e. not all degrees of freedom (DoF) are observable.

This is even drastically worse when using a single monocular camera as the scale is also unknown. For all this, one must incorporate some *a-priori* knowledge into the problem in order to confine the possible solutions into a reduced set of solutions that correctly represents the real nature of the environment. In this paper, we propose the following assumptions to constrain our reconstruction problem:

- 1) **Local isometry:** we assume that the vicinity of a surface point follows an isometric model, that is local distances are preserved.
- 2) **Smooth deformation:** we consider that points that are close in space must undergo similar deformations.

- 3) **Slow deformation:** deformations are assumed to happen slowly over time.
- 4) **Camera motion is faster than deformation:** finally, we assume that camera can move faster than deformations, so we attribute rigid motions to the camera, computing deformations as small as possible.

Assumption (1) enables us to use a photometric feature tracker defining a small neighbourhood around each tracked point that is assumed to be locally rigid. This allows us to take an approach similar to [18] to perform short term data association. Assumption (2) introduces local constraints in the deformations observed without imposing a global deformation or surface model. This effectively makes our system general enough to model any environment. Assumption (3) allows us to impose temporal continuity in the position of surface points, reducing the effect of image and data association noise. Finally, Assumption (4) is the one that allows us to separate camera motion and environment deformation. In all SLAM systems, the sensor provides relative information, and as a result, the absolute pose of camera and environment is not observable. In rigid SLAM this is simply addressed by choosing an arbitrary global pose, for example the first camera pose is chosen to be zero. In deformable SLAM this is not enough as a camera motion is indistinguishable from a hypothetical case where all the environment moves rigidly, what is called the *floating map ambiguity* in [17]. This assumption allows us to use regularizers that penalize deformation over camera motion. In that way, the rigid part of the relative motion between environment and camera will be attributed to camera motion, obtaining deformations as small as possible.

It is important to note that none of the above assumptions impose global constraints over the surface topology, smoothness, or deformation, allowing us to model generic deformations and environments.

### B. Data Association

Our previous experience in monocular deformable SLAM [6] has proven that an accurate data association is crucial in order to reach good accuracy and robustness. Indeed other works have shown the potential of direct methods in this task like [2] in which the photometric term allows to get as a byproduct of the tracking the feature associations. This is done by imposing a global rigid transformation to all points as it is assumed that the environment is stationary. However this is far to be true in deformable SLAM. Indeed one can not impose any global constraint to the data association step as it is easily violated by deformations.

For that, we propose to perform photometric data association with Shi-Tomasi features [19] prior the camera pose and deformation estimation using the modified multi-scale Lucas-Kanade algorithm proposed in [6]:

$$\arg \min_{\mathbf{d}, \alpha, \beta} \sum_{\mathbf{v} \in P(\mathbf{u})} (I^0(\mathbf{v}) - \alpha I^t(\mathbf{v} + \mathbf{d}) - \beta)^2 \quad (1)$$

where  $P(\mathbf{u})$  is a small pixel patch centered at the keypoint  $\mathbf{u}$ .  $I^0$  is the first frame, where the points are initialized, and  $I^t$

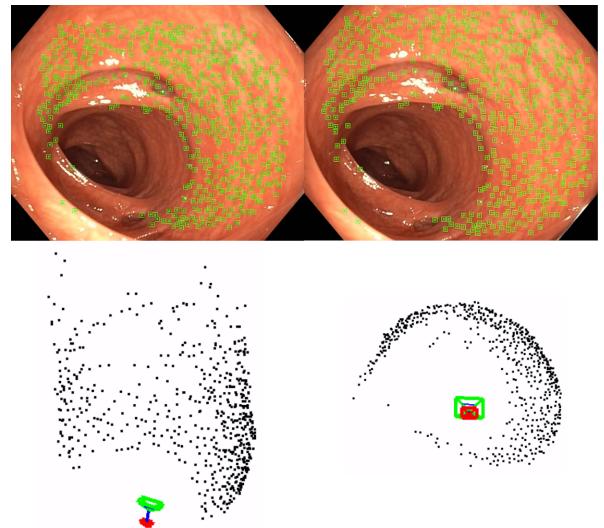


Fig. 2: Top: Two images separated by 3 frames from our EndoMapper dataset with tracked features. Bottom: map initialized from those tracks

is the current frame in time  $t$ . These patches are updated every 5 images to account for big scale changes or rotations. This algorithm has been proven to achieve excellent results when tracking image features in short time steps even in the presence of deformations or local illumination changes (Fig. 2). The key of its performance lies in using no global model: each point can move freely with respect to the others. Also a local illumination invariance is achieved by computing local gain  $\alpha$  and bias  $\beta$  terms for each point.

In order to remove any possible outlier track, we compute the *Structural Similarity Index* (SSIM) [20] between the reference  $x$  and tracked  $y$  pixel patches to identify any outlier track:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2)$$

where  $\mu_x$  and  $\sigma_x$  are the mean and covariance of the pixel patch,  $\sigma_{xy}$  is the cross covariance between both patches and  $C_1$  and  $C_2$  are constant values to avoid instability when means and covariances approach zero. This has been proven to be a good similarity metric for small pixel windows as it combines in a same metric a luminance, contrast and structure comparison.

### C. Monocular map initialization

Initializing a map from monocular images in rigid environments is well known in Structure from Motion (SfM). In deforming environments, Non-Rigid Structure from Motion (NRSfM) techniques can be used [7]. However, they require, for example in the map initialization, assumptions such as a smooth scene surface with planar topology, which are not met in real colonoscopies (see Fig. 1 and 2).

We propose to exploit assumption (4) using two close frames in which the environment can be considered quasi-rigid, and most image innovation can be attributed to the camera motion. This allows to apply SfM to obtain a first estimation of the map as if it is rigid and treat any deformation as small noise.

Ideally, the method should to be independent of the camera model either pinhole or fish-eye. We propose to initialize the monocular map by computing the Essential matrix between 2 close frames using as input normalized projective rays from features in the images. Our proposed initialization algorithm goes through the following steps:

- 1) Extract Shi-Tomasi features evenly distributed in the reference frame  $I^0$  and track them in the current frame  $I^t$  using the Lucas-Kanade optical flow algorithm. Unproject the matched features into normalized rays  $\mathbf{x}_i^0$  and  $\mathbf{x}_i^t$  and using the camera model unprojection function.
- 2) Compute an Essential matrix that relates poses of the two frames:

$$\mathbf{x}_i^{tT} \mathbf{E} \mathbf{x}_i^0 = 0 \quad (3)$$

This is done inside a RANSAC scheme to reduce the influence of outliers coming from the data association.

- 3) Recover the relative camera motion  $\mathbf{T}_{C^t C^0}$  from  $\mathbf{E}$ . This will yield to 4 motion hypothesis (2 rotations and 2 translations). We are using close frames to initialize, hence the camera rotation should be small so we can safely select the smallest rotation to solve the rotation ambiguity. Finally we disambiguate the translation component by selecting the one that yields to the highest number of points in front of both cameras.
- 4) Reconstruct environment using the camera motion recovered. For that, we use the Inverse Depth Weighted Midpoint [21] to triangulate tracked features as it provides low 3D-2D errors in low parallax scenarios.

#### D. Camera pose prediction

To encode assumption (4), we first estimate a preliminary camera pose  $\mathbf{T}_{C^t W}$  for time  $t$  prior estimating any deformation. We assume that the camera follows a physical model of constant velocity. This provides us with an initial guess of the camera pose that will then be refined using Non Linear Least Squares (NLLS) using a reprojection error with the environment geometry observed in the previous frame  $t - 1$ . This can be seen as a way to attribute to a camera motion most of the image innovation seen. Note that this is not the final pose we compute but just a seed for our global optimization for the deformations and camera pose detailed in the next section.

#### E. Tracking camera pose and deformation

Our goal is, given some feature matches in the current frame  $\mathbf{u}_i^t$  and the 3D reconstruction in the previous temporal instant  $\mathbf{X}_i^{t-1}$ , to estimate the current camera pose  $\mathbf{T}_{C^t W}$  and the deformation  $\delta_i^t$  of each point such as the current scene can be estimated as  $\mathbf{X}_i^t = \mathbf{X}_i^{t-1} + \delta_i^t$ .

For that we introduce a reprojection data term  $E_{i,rep}^t$  along with 2 regularizers  $E_{i,spa}^t$  and  $E_{i,tmp}^t$  to constrain the deformations in our total global cost function,  $\mathcal{E}^t$  for time  $t$ , defined by:

$$\mathcal{E}^t = \sum_{i \in \mathcal{P}} E_{i,rep}^t + \lambda_{spa} E_{i,spa}^t + \lambda_{tmp} E_{i,tmp}^t \quad (4)$$

where  $\mathcal{P}$  represents the set of points being observed in the current frame. Our global problem can be solved using Non-Linear Squares optimization such us:

$$\mathbf{T}_{C^t W}, \delta_i^t = \arg \min_{\mathbf{T}_{C^t W}, \delta_i^t} \mathcal{E}^t \quad (5)$$

Next we define the terms of our cost function  $\mathcal{E}^t$ .

1) *Reprojection term*: we obtain feature matches  $\mathbf{u}_i^t$  in the current frame with the modified Lucas-Kanade algorithm presented in [6], computing on this way the reprojection error as it follows:

$$E_{i,rep}^t = \rho(\|\mathbf{u}_i^t - \hat{\mathbf{u}}_i^t\|_{\Sigma_{rep}}^2) \quad (6)$$

where  $\rho$  is the Hubber robust cost,  $\hat{\mathbf{u}}_i^t$  and  $\mathbf{u}_i^t$  are respectively the match of feature  $i$  in the current image  $I_t$  and its projection given by:

$$\hat{\mathbf{u}}_i^t = \Pi(\mathbf{T}_{C^t C^0}(\Pi^{-1}(\mathbf{u}_i^0, d_i) + \mathbf{X}_i^{t-1} + \delta_i^t)) \quad (7)$$

The accuracy of indirect methods is limited by the feature detector resolution (typically no less than 1 pixel). However matches obtained with photometric methods have subpixel accuracy boosting on this way the accuracy of our reprojection term while keeping its nice convergence basin.

2) *Spatial regularizer*: Following [11] we encode assumption (2) with a regularizer that constrains deformations locally so they are spatially smooth:

$$E_{i,spa}^t = \sum_{j \in \mathcal{G}(i)} \rho(\|w_{ij}^t(\delta_i^t - \delta_j^t)\|_{\Sigma_{spa}}^2) \quad (8)$$

Here  $\mathcal{G}$  represents a weighted graph that encodes related points whose deformations should be regularized together. The weight in  $\mathcal{G}$  of two connected points  $i$  and  $j$  is  $w_{ij}^t$  which depends on the Euclidean distance between both points at the immediately previous time instant  $t - 1$  and is computed according to the following formula:

$$w_{ij}^t = \exp\left(\frac{-\|\mathbf{X}_i^{t-1} - \mathbf{X}_j^{t-1}\|^2}{2\sigma^2}\right) \quad (9)$$

where  $\sigma$  is a radial basis weight that controls the influence radius of each point. This regularizer is crucial as it enforces as rigid-as-possible deformations and contributes towards a global consistency of the deformations.

3) *Temporal regularizer*: Finally we add a temporal regularizer on the deformations to represent that they occur slowly over time (assumption (3)):

$$E_{i,\text{tmp}}^t = \rho(\|\delta_i^t\|_{\Sigma_{\text{tmp}}}^2) \quad (10)$$

This regularizer also interacts with assumption (4) as it penalizes big deformations that could be explained with a camera motion.

#### IV. EXPERIMENTS

We evaluate our method in a Minimal Invasive Surgery sequences, more specifically in colonoscopies. This kind of sequences pose big challenges as they exhibit continuous deformations, poor textures and harsh illumination conditions. We provide quantitative results in photorealistic synthetic data and qualitative experiments with in-vivo human colonoscopies. For comparison purposes we also test our method in the Hamlyn dataset using its stereo setup to evaluate our reconstructions against other state-of-the-art methods. A summary of our main results can be seen in Fig 3.

##### A. Implementation details

We implement the monocular map initialization, camera pose and deformation estimation in C++. For non Linear Squares optimization we use the Levenberg-Marquart algorithm implemented in the g2o library [22]. For feature extraction and matching we implement our own Shi-Tomasi feature extractor and Lucas-Kanade tracker (Eq. 1). We set a threshold of 0.8 for the SSIM score (Eq. 2) to detect and reject spurious feature tracks. Regarding the optimization, we set  $\Sigma_{\text{rep}}$  to 1 pixel,  $\Sigma_{\text{spa}}$  and  $\Sigma_{\text{tmp}}$  to 10 mm. Since  $\Sigma_{\text{spa}}$  and  $\Sigma_{\text{tmp}}$  correctly scales the  $E_{\text{spa}}$  and  $E_{\text{tmp}}$  terms, we set their respective  $\lambda$  to 1. Finally, for the Hubber cost threshold we use the 95 percentile of  $\chi^2$ , with 2 DoF for  $E_{\text{rep}}$  and with 3 DoF for  $E_{\text{spa}}$  and  $E_{\text{tmp}}$ .

Regarding the regularization graph  $\mathcal{G}$ , for each point we only add regularization terms with its  $K = 20$  closest points in 3D with  $\sigma$  set to 15 mm when initializing the map with the monocular camera and 55 mm when using the stereo to get the first map reconstruction. This is done to ensure that a points is always regularized and at the same time ignoring points that have little influence with the current point to reduce the computational burden.

##### B. Simulated Colon dataset

We use the VR-Caps [8] to generate photorealistic synthetic image sequences of a 3D colon model obtained from a Computed Tomography. Since this is a simulation, we have access full to camera pose and 3D scene ground truth. Indeed, we can generate sequences with different camera trajectories and degrees of deformation enabling us to test each one of the components of our system individually.

For evaluation purposes, we simulate an insertion maneuver (Fig. 3b) with different degrees of deformation. We model the deformations via a sine wave propagating along the simulated colon. We apply this deformations to

$A$ (mm)	$\omega$ (rad/s)	0	2.5	5
0	1.15	-	-	-
2.5	-	1.77	1.70	-
5	-	1.84	3.65	-
10	-	2.27	4.57	-

TABLE I: Reconstruction RMSE (mm) in simulated colonoscopies [23] for different deformation types

the  $y$  coordinate of the point surfaces simulating peristaltic movements according to the following formula:

$$V_y^t = V_y^0 + A \sin(\omega t + V_x^0 + V_y^0 + V_z^0) \quad (11)$$

where  $V_x^0$ ,  $V_y^0$  and  $V_z^0$  are the coordinates of the surface point at rest. We can control the magnitude and velocity of the deformations according to the parameters  $A$  and  $\omega$  respectively. Table I shows the reconstruction accuracy of our system in the simulated sequence with different deformation velocities and amplitudes. The error shown is the Root Mean Square Error (RMSE) of the reconstructed points for all frames according to:

$$e_{\text{rms}} = \sqrt{\frac{\sum_i \|s^t \hat{\mathbf{X}}_i^t - \mathbf{X}_i^{t,\text{gt}}\|^2}{n}} \quad (12)$$

Since this is a full monocular formulation, we find, for each frame, an optimal scale factor,  $s^t$ , to align our reconstructions with the ground truth.

Results show that our formulation can reach nice reconstruction error around 2-3 mm even though in presence of deformations. One interesting result is that our system is more sensitive to deformation velocities than the magnitude itself being aligned with assumption (3).

##### C. Hamlyn dataset

We also test our formulation in real endoscopic sequences. For that purpose, we use sequences 20 (Fig. 3c) and 21 (Fig. 3d) of the Hamlyn dataset [10]. Sequence 20 (from frame #750) corresponds to abdominal exploration with slow deformation. Sequence 21 (also from frame #750) images a liver with 2 lobes each of them moving on its own. This can be considered as an articulated motion. In both sequences, surface texture is poor and illumination conditions are unfavorable. This dataset is recorded with a stereo endoscope, allowing us to estimate environment groundtruth from the disparity observed by the stereo sensor.

We evaluate our formulation in 2 setups (Table II) for comparison purposes. In the first setup, we initialize our system with the first stereo images and perform monocular tracking, in order to allow comparison with previous methods ORB-SLAM [1], SD-DefSLAM [6] and Direct and Sparse Deformable Tracking (DSDT) [17]. Since we are initializing from the stereo images, we do not perform any scale alignment when computing the RMSE. We achieve

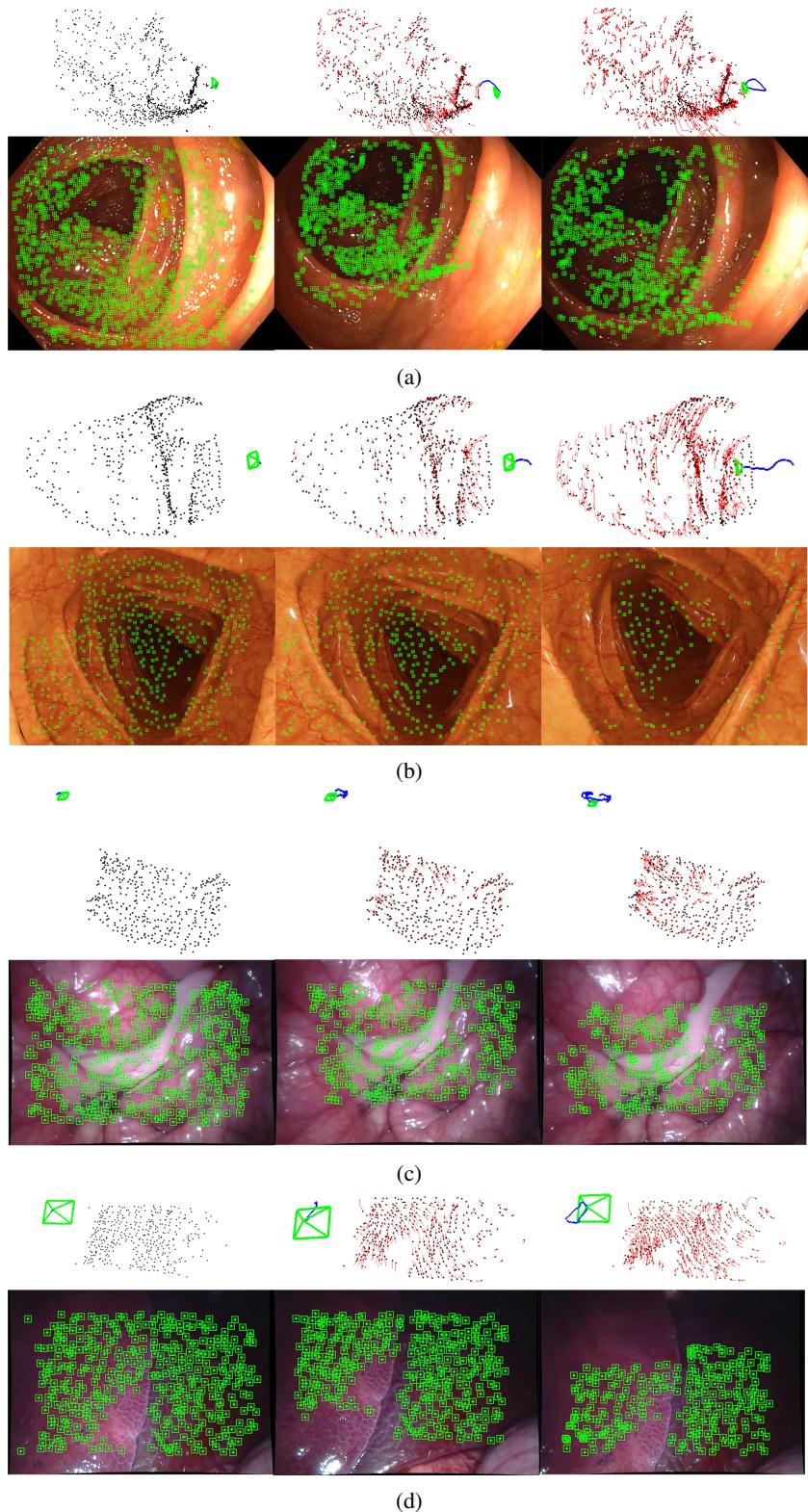


Fig. 3: Results of our algorithm for different sequences. Per each sequence, in columns results after the initial middle and final frame. Two rows per sequence. The first row displays the 3D reconstruction: black points are the undeformed map, red lines are the map point deformation trajectories, in blue the camera trajectory. The second row the RGB frames with the tracked features in green. From top to bottom: (a) EndoMapper real in-vivo sequence, (b) Simulated sequence, (c) Hamlyn 20 sequence and (d) Hamlyn 21 sequence. All datasets have been processed using only monocular images.

		Stereo Initialization				Monocular Init.
		ORB-SLAM3 [4]	SD-DefSLAM [6]	DSDT [17]	Ours	Ours
20	RMSE	1.37	4.68	2.9	1.48	2.79
	# Fr.	220	252	500	350	350
21	RMSE	-	6.19	1.3	1.55	3.31
	# Fr.	-	323	300	300	300

TABLE II: Comparison with previous methods in sequences 20 and 21 from Hamlyn Dataset as shown in [17]. We report reconstruction RMSE (mm) and number of frames processed.

competitive results regarding reconstruction error, obtaining a consistent error around 1.5 mm. Since we do not impose any restriction on the surface topology or in the deformations, we achieve a significantly smaller error compared with SD-DefSLAM. This is specially clear in sequence 21 with the 2 lobes moving independently what limits the accuracy of SD-DefSLAM. The comparison with DSDT suggests that our regularizers are versatile, we are able to code better the spatial smoothness of sequence 20, achieving a lower error, while still being competitive in hard discontinuity of sequence 21. DSDT is able to keep the track longer because, in contrast with DSDT, our method still does not implement any policy to recover points lost during tracking.

The second setup uses the full monocular pipeline including our monocular initialization (Sec. III-C) computing the RMSE after a per frame scale correction (Eq. 12). In this scenario, our system reaches errors around 2.8-3.3 mm which is aligned with the errors obtained in the simulation dataset under significant deformations. The increase in error compared with the stereo setup is due to the quality of the map initialization that no longer relies on a perfect stereo initialization.

Also it is important to note that in these sequences the surfaces shape and deformations observed are completely different from the ones seen in the simulation dataset proving that we can model general surface shapes and deformations.

#### D. Real endoscopy sequences

We provide qualitative results in real in-vivo human colonoscopy sequences from the EndoMapper project [9]. These sequences display the big challenges real colonoscopies pose, such as deformation, little to no texture in the images, lighting conditions varying from frame to frame, reflections and fish-eye optics (fig. 1.)

In this case, there is no ground truth to compare with, because the dataset just records standard monocular endoscope procedures. For this reason, we only provide qualitative results. Fig. 2 displays how we are able to initialize maps with high density of points form quite close (3 frames apart) frames capturing the tubular topology of the colon. In Figures 1 and 3a, it can be seem how our algorithm is able to capture the scene deformation and the endoscope trajectory, being able to track map points for more than 30 frames in two examples of real colonoscopies of two different patients.

## V. CONCLUSIONS

In this work we have presented an approach for monocular camera tracking and deformation estimation without assumptions on the environment shape or topology. Instead, we successfully encode with simple regularizers the assumptions about the type of deformations that are common in endoscopy. Compared with the state of the art, and our method, including map initialization, is applicable in a much wider range of shape and topologies, like colonoscopies, while having similar accuracy in more standard almost-planar scenarios.

The presented monocular initialization and tracking contributes to make real a fully deformable SLAM system. The deformable mapping to expand the map as the camera explores new regions is closer after our contribution, being a promising venue of future work in the short term. In the mid-term, a multi-map deformable SLAM offers a profitable for future work because will be able to cope with occlusions and tracking losses prevalent in real colonoscopies.

## REFERENCES

- [1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, “ORB-SLAM: a versatile and accurate monocular SLAM system,” *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [2] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.
- [3] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, “SVO: Semidirect visual odometry for monocular and multicamera systems,” *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2016.
- [4] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM,” *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.
- [5] J. Zubizarreta, I. Aguinaga, and J. M. M. Montiel, “Direct sparse mapping,” *IEEE Transactions on Robotics*, vol. 36, no. 4, pp. 1363–1370, 2020.
- [6] J. J. Gómez-Rodríguez, J. Lamarca, J. Morlana, J. D. Tardós, and J. M. Montiel, “SD-DefSLAM: Semi-direct monocular SLAM for deformable and intracorporeal scenes,” in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2021, pp. 5170–5177.
- [7] J. Lamarca, S. Parashar, A. Bartoli, and J. Montiel, “DefSLAM: Tracking and mapping of deforming scenes from monocular sequences,” *IEEE Transactions on robotics*, vol. 37, no. 1, pp. 291–303, 2020.
- [8] K. Incetan, I. O. Celik, A. Obeid, G. I. Gokceler, K. B. Ozyoruk, Y. Almaliooglu, R. J. Chen, F. Mahmood, H. Gilbert, N. J. Durr *et al.*, “Vr-caps: a virtual environment for capsule endoscopy,” *Medical image analysis*, vol. 70, p. 101990, 2021.
- [9] “Endomapper project,” <https://sites.google.com/unizar.es/endomapper/home>, accessed: 2022-02-28.

- [10] P. Mountney, D. Stoyanov, and G.-Z. Yang, “Three-dimensional tissue deformation recovery and tracking,” *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 14–24, 2010.
- [11] R. A. Newcombe, D. Fox, and S. M. Seitz, “DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 343–352.
- [12] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, “VolumeDeform: Real-time volumetric non-rigid reconstruction,” in *European Conference on Computer Vision*. Springer, 2016, pp. 362–379.
- [13] M. Slavcheva, M. Baust, and S. Ilic, “SobolevFusion: 3D reconstruction of scenes undergoing free non-rigid motion,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2646–2655.
- [14] ——, “Variational level set evolution for non-rigid 3D reconstruction from a single depth camera,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2838–2850, 2020.
- [15] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake, “Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery,” *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 155–162, 2017.
- [16] H. Zhou and J. Jayender, “EMDQ-SLAM: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 331–340.
- [17] J. Lamarca, J. J. G. Rodriguez, J. D. Tardos, and J. Montiel, “Direct and sparse deformable tracking,” *arXiv preprint arXiv:2109.07370*, 2021.
- [18] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of the 7th International Joint conference on Artificial Intelligence*, vol. 2, 1981, p. 674–679.
- [19] J. Shi and C. Tomasi, “Good features to track,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.
- [20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [21] S. H. Lee and J. Civera, “Triangulation: why optimize?” *arXiv preprint arXiv:1907.11917*, 2019.
- [22] G. Grisetti, R. Kümmerle, H. Strasdat, and K. Konolige, “g2o: A general framework for (hyper) graph optimization,” in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 9–13.
- [23] K. B. Ozyoruk, G. I. Gokceler, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, H. Sahin, H. Araujo, H. Alexandrino, N. J. Durr, H. B. Gilbert, and M. Turan, “Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos: Endo-sfmllearner,” 2020.