2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

DefSLAM: Tracking and Mapping of Deforming Scenes From Monocular Sequences

Jose Lamarca, Shaifali Parashar, Adrien Bartoli, and J. M. M. Montiel, *Member, IEEE*

***Abstract*—Monocular simultaneous localization and mapping (SLAM) algorithms perform robustly when observing rigid scenes; however, they fail when the observed scene deforms, for example, in medical endoscopy applications. In this article, we present DefSLAM, the first monocular SLAM capable of operating in deforming scenes in real time. Our approach intertwines Shape-from-Template (SfT) and Non-Rigid Structure-from-Motion (NRSfM) techniques to deal with the exploratory sequences typical of SLAM. A deformation tracking thread recovers the pose of the camera and the deformation of the observed map, at frame rate, by means of SfT processing a template that models the scene shape-at-rest. A deformation mapping thread runs in parallel with the tracking to update the template, at keyframe rate, by means of an isometric NRSfM processing a batch of full perspective keyframes. In our experiments, DefSLAM processes close-up sequences of deforming scenes, both in a laboratory-controlled experiment and in medical endoscopy sequences, producing accurate 3-D models of the scene with respect to the moving camera.**

***Index Terms*—Deformable simultaneous localization and mapping, minimal invasive surgery, real-time systems, robustness, simultaneous localization and mapping, SLAM, strain, surgery, surgical vision, three-dimensional displays.**

I. Introduction

THE goal of visual simultaneous localization and mapping (SLAM) algorithms is to locate a visual sensor in an uncertain map which is being estimated simultaneously. The typical use case in SLAM includes exploratory trajectories where the camera images a scene without previous information of the structure observed. Using a monocular sensor, visual SLAM has to process several images rendering enough parallax to recover the map for the new scene region with respect to the camera. Once the map is available, the camera can be localized with respect to this map from just one image as long as the camera does not move to unexplored areas. The rigidity assumption constrains the problem significantly, and it is intensively exploited by state-of-the-art monocular SLAM systems [1]–[3].

However, the rigidity assumption rends invalid in applications where the deformation is predominant. To this end, we introduce DefSLAM, a calibrated monocular and deformable SLAM system which can perform in deforming, i.e., nonrigid, environments. A relevant use case is medical endoscopy, where monocular visual SLAM is a crucial tool for augmented reality and autonomous medical robotics.

In the literature, nonrigid monocular scenes have been handled by Non-Rigid Structure-from-Motion (NRSfM) [4]–[8] and Shape-from-Template (SfT) [9]–[12] methods. NRSfM methods are able to recover the evolution of the 3-D scenes nonrigid deformations from a set of monocular images, after a computationally demanding batch processing of the images. In contrast, SfT recovers the 3-D deformation from a single image, at a low computational cost but needs a template. The template is a 3-D textured model describing the shape at rest of the scene. DefSLAM framework combines the advantages of the two classes of nonrigid monocular methods. We propose a parallel algorithm composed of a *deformation tracking* thread as the front-end running SfT at frame rate, and a *deformation mapping* thread as the back-end running NRSfM to compute the SfT template at a slower keyframe rate.

Fig. 1 shows DefSLAM processing a sequence where the camera is being located with respect to a deforming kerchief being mapped simultaneously using images from a monocular sensor from partial observations of different regions of the kerchief. The *deformation tracking* thread recovers the camera pose and the deformation of the map at frame rate. It uses a template for the viewed part of the map to recover the map points deformation by minimizing a combination of reprojection error and deformation energy for each frame. The *deformation mapping* thread initializes and refines map estimates, and extends the map when new regions are visited. It processes just a selection of frames—keyframes—imaging the same region to define the shape-at-rest of the template used by the *deformation tracking* thread to process the subsequent frames.

We validated our DefSLAM algorithm in monocular sequences that include exploratory trajectories observing deforming scenes. We evaluate DefSLAM on new waving mandala kerchief dataset which we created and an *in vivo* medical endoscopy Hamlyn dataset [13]. To make some comparison, we have resorted to systems with a different configuration than

Manuscript received May 29, 2020; accepted July 31, 2020. Date of publication September 21, 2020; date of current version February 4, 2021. This work was supported in part by the European Unions Horizon 2020 research and innovation programme under Grant 863146, in part by the PGC2018-096367-B-I0 MCIU/AEI/FEDER, UE, in part by the Spanish Agencia estatal de investigación DPI2017-91104-EXP, and in part by the MINECO scholarship BES-2016-078678. *(Corresponding author: Jose Lamarca.)*

Jose Lamarca and J. M. M. Montiel are with the Instituto de investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, 50009 Zaragoza, Spain (e-mail: jlamarca@unizar.es; josemari@unizar.es).

Shaifali Parashar and Adrien Bartoli are with the Institut Pascal – UMR 6602 – CNRS/UCA/CHU, F-63000 Clermont-Ferrand, France (e-mail: shaifali.parashar@gmail.com; Adrien.Bartoli@gmail.com).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2020.3020739

978-1-7281-9077-8/21/$31.00 ©2021 IEEE

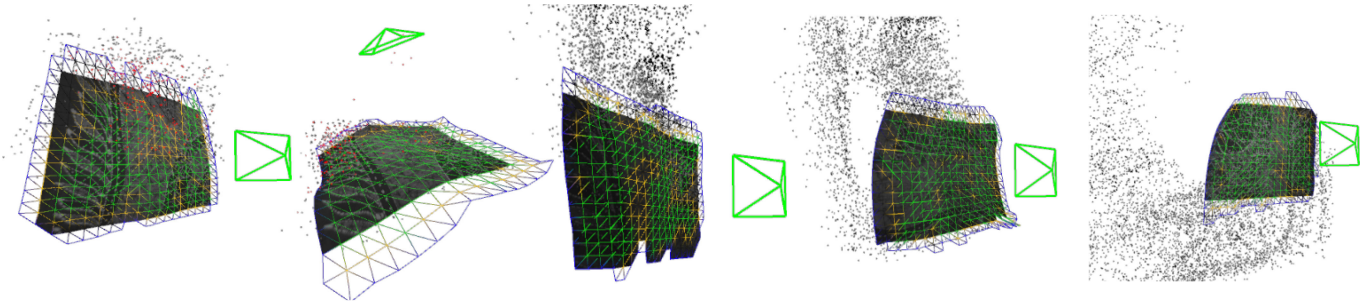


Fig. 1. Real-time reconstruction of a deforming scene with DefSLAM. The mandala kerchief deforms while the camera moves. DefSLAM locates the camera shown as a green frustum, while recovering the deformation of the kerchief using a template of the same. The estimated 3-D deformable map is expanded when new regions are explored by reestimating new templates. The map is composed of sparse 3-D points, in black, and a template as triangular mesh, viewed part in green.

ours. We compare our results with the state-of-the-art rigid monocular ORBSLAM [3] to display the DefSLAM unique capability to SLAM deforming scenes. We also compared with MISSLAM [14], the closest in the literature offering SLAM accuracy results in medical deformable scenes, despite being a stereo in contrast to our monocular system. These experiments validate the unprecedented ability of DefSLAM to accurately code the structure of the scene in rigid and deformable scenarios, including medical cases.

II. Related Work

A. SLAM

Deformable Visual SLAM: The deformable SLAM methods in the literature rely on sensors providing depth information, i.e., RGB-D or stereo sensors. DynamicFusion [15] is a seminal work in deformable VSLAM with an RGB-D camera. It fuses the frame-by-frame depth information into a canonical shape, i.e., a shape-at-rest, that incrementally maps the entire scene after an exploratory trajectory of partial observations. This canonical shape is deformed to the current keyframe with the as-rigid-as-possible deformation model [16]. In [17], the quality of the deformation is improved by including the photometric error in the optimization. In [18], the volumetric representation is substituted by surfels to improve the efficiency of the algorithm. These methods recover the whole canonical shape deformation which is usually small. This technique is not scalable to bigger shapes like exploratory scenes in endoscopy. Gotardo and Martinez [19] propose to use an embedded deformation model [20] instead of as-rigid-as-possible because it better preserves the local details under the deformation. In [14], the system is enhanced with the tracking of a rigid system ORBSLAM [3] to achieve better tracks and more robust deformable SLAM for medical endoscopy exploration. In any case, all these algorithms optimize the whole map each time and thus scale poorly with the size of the map. We aim similar SLAM capabilities in deformable scenes, but in the challenging monocular case. In addition, our approach only optimizes the observed map zone achieving good scalability with respect to the size of the map, being able to be run on the CPU.

Rigid Visual SLAM: Monocular rigid VSLAM is a mature field. The current state-of-the-art monocular rigid VSLAM methods such as [1], [3] provide accurate, robust, and fast results in robotic scenes. Some works have attempted to apply rigid methods in *in vivo* medical quasi-rigid scenes. Innmann *et al.* [21] propose an EKF-SLAM algorithm, and Klein and Murray [22] get dense maps based on [3]. Kummerle *et al.* [23] use a rigid SLAM system to locate the camera in arthroscopic images. All of these methods assume that the deformation is negligible and hence that a purely rigid SLAM system is able to survive just by excluding from the map any deformed scene region. We aim to achieve a similar performance, but in scenarios where deformation is predominant, more specifically, real-time operation and capability to handle sequences of close-ups corresponding to exploratory trajectories.

B. Nonrigid Monocular Techniques

The methods in the literature which aim to recover the structure of a nonrigid scene from monocular sequences are SfT and NRSfM.

Shape-From-Template: SfT methods recover the deformed shape of an object from a monocular image and the object's textured 3-D shape at rest. This textured shape-at-rest of the object is the so-called *template*. These methods associate a deformation model with this template to recover the deformed shape. The main difference between these methods is the definition of the deformation model. We distinguish between analytic and energy-based methods. Among the analytic solutions, we focus on the isometric deformation which assumes that the geodesic distance between points in the surface is preserved. Isometry for SfT has proven to be well-posed and it quickly evolved to stable and real-time solutions [9], [24], [25]. Energy-based methods [10]–[12], [26] jointly minimize the shape energy with respect to the shape-at-rest and the reprojection error for the image correspondences. These optimization methods are well suited to implement sequential data association with robust kernels to deal with outliers.

Orthographic Nonrigid Structure-From-Motion: The earliest nonrigid monocular techniques are NRSfM. These methods were formulated using statistical models, first proposed in [27]. This work gave rise to a family of methods [28]–[30] which used a low-dimensional basis model to obtain the configuration of the 3-D points from the images of a sequence. They exploited spatial regularizers [28], [31], temporal regularizers

[32], and spatio-temporal regularizers [33]–[35]. These methods may handle small surface deformations or articulated objects, but they usually fail with very large deformations. They use an orthographic camera model which is an approximation only valid when the scene is distant from the camera; this is a strong assumption invalid in many applications.

Perspective Nonrigid Structure-From-Motion: Real use cases need the more accurate perspective camera model. It is able to model the close-up sequences typical in SLAM, especially in medical endoscopy. The isometry assumption, first proposed in SfT methods, has also produced excellent results in NRSfM [4]–[8]. It brought not only improvements in terms of accuracy, but also the ability to handle perspective cameras. Ref. [6] is a local method able to handle naturally occlusions and missing data also usual in many applications.

Our Approach: We propose the first visual SLAM system capable of working with deforming monocular sequences. We propose a *deformation tracking* thread based on [10], which uses a precomputed template to recover the camera pose and the deformation of the scene. We also propose a *deformation mapping* thread which extends the map and estimates the shape-at-rest of the template in new explored zones by means of the isometric NRSfM proposed by [6]. Our contribution is a new iterative scheme for the optimization in [6] that allows to calculate and refine the solutions incrementally at keyframe rate. Both for the deformation mapping and tracking, we only optimize the part of the template observed having a runtime independent of the size of the map in exploratory sequences.

We also propose a sequential active matching that exploits the already available SLAM map to boost the data association performance. Our final contribution is to integrate in the deformation mapping an alignment between surfaces to build a global map, extending alignment as proposed in [14], [15], [18] to the monocular case.

The proposed deformation tracking and mapping algorithms can run in parallel, in a similar way to the state-of-the-art rigid SLAM methods [1]–[3] to achieve real-time performances.

III. DefSLAM System Overview

DefSLAM recovers the structure of the scene, its deformation, and the camera pose. It is composed of three main components:

1) *The map:* The map represents the structure of the scene reconstructed by DefSLAM as a set of 3-D map points. The map is deformable and the position of the map points evolves along the sequence. Each map point j is represented by its position $\mathbf{X}_j^t$ for each processed frame t.
   We save some selected frames in the map called keyframes. We refer to the keyframes in which a map point is initialized as anchor keyframes. After each new keyframe processing, one of the anchor keyframes is selected as the reference keyframe. The reference keyframe defines the template used by the deformation tracking to process the new incoming frames.
2) *The deformation tracking thread:* This thread is the front-end of the system and runs at frame rate. It uses SfT to estimate the position of the map points $\mathbf{X}_j^t$ and the camera pose $\mathbf{T}_{tw}$ for each frame t. We embed the map points into the template $\mathcal{T}_k$ to compute their position. The shape-at-rest of the template $\mathcal{T}_k$ is the surface $\mathcal{S}_k$ observed in the reference keyframe k.
3) *The deformation mapping thread:* This thread is the back-end of the system and runs at keyframe rate. It uses NRSfM to estimate the surface $\mathcal{S}_k$ observed in the keyframe k.

Notation: We use calligraphic letters for sets of geometrical entities in the deforming scene, e.g., $\mathcal{X}$ for the set of all map points. Bold letters represent matrices and vectors. Scalars are represented in italics. The indexes t represent the frames and $\mathbf{T}_{tw}$ the pose of the frame at instant *t*. Superindexes represent the temporal instant of the estimation. The index j represents the map points, n the nodes, and e the edges of the mesh describing the template surface.

IV. Deformation Tracking

Deformation tracking recovers the camera pose $\mathbf{T}_{tw}$ and the shape of the template $\mathcal{T}_k^t$ in the frame t by jointly minimizing reprojection error and deformation energy. $\mathcal{T}_k$ is the surface reconstructed in the reference keyframe k. The tracking algorithm is composed of three stages: data association, camera pose estimation, template deformation, and new keyframe selection. Next, the template structure, the camera model, and the three steps of the algorithm are detailed.

A. Template

The template is a surface parametrized with a 3-D triangular mesh. It is composed of a set of planar triangular facets $\mathcal{F}$, defined by a set of nodes $\mathcal{V}$, and connected by a set of edges $\mathcal{E}$. The deformation of the map at frame t is defined through the pose of the nodes of the template $\mathcal{T}_k^t$. The facet $f \in \mathcal{F}$ at frame t is defined by the pose of its three nodes $V_{f_j}^t = \{V_{f,h}^t\},\ \ h = \{1,2,3\}$. The map points observed in the keyframe k are embedded in the facets of the mesh. The position of a map point $\mathbf{X}_j^t \in \mathcal{X}$ in frame t is defined with its barycentric coordinates, $\boldsymbol{b}_j = [b_{j,1}, b_{j,2}, b_{j,3}]^\top$, with respect to the position of the nodes of the face f_j:

$$\boldsymbol{X}_j^t = \sum_{h=1}^{3} b_{j,h} \boldsymbol{V}_{f_j,h}^t \text{ s.t. } b_{j,1} + b_{j,2} + b_{j,3} = 1. \tag{1}$$

B. Camera Model

We use the calibrated pinhole model. The projection of the 3-D point j, $\boldsymbol{X}_j^t \in \mathcal{X}_k^t$ in the frame t by a camera located at $\boldsymbol{T}_{tw}$ is modeled by the projection function $\pi : [\mathrm{SE}(3), \mathbb{R}^3] \to \mathbb{R}^2$:

$$\pi\left(\boldsymbol{T}_{tw}, \boldsymbol{X}_j^t\right) = \begin{bmatrix} f_x \frac{X_j^t}{Z_j^t} + C_x \\ f_y \frac{Y_j^t}{Z_j^t} + C_y \end{bmatrix}$$

$$\text{where } \begin{bmatrix} X_j^t & Y_j^t & Z_j^t \end{bmatrix}^\top = \boldsymbol{R}_{tw} \boldsymbol{X}_j^t + \boldsymbol{t}_{tw}. \tag{2}$$

$\boldsymbol{R}_{tw} \in SO(3)$ and $\boldsymbol{t}_{tw} \in \mathbb{R}^3$ are, respectively, the rotation and the translation of the transformation $\boldsymbol{T}_{tw}$. $\{f_x, f_y, C_x, C_y\}$ are

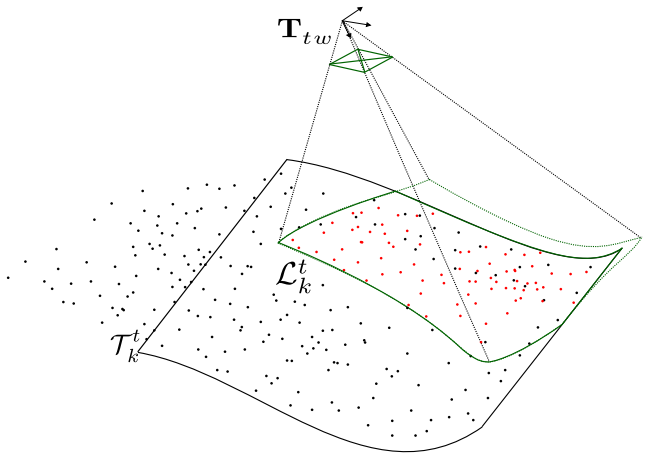


Fig. 2. Deformation tracking: estimating camera pose and deformation of the viewed map. $\mathcal{T}_k^t$ is the map shape in the frame t, $\mathcal{L}_k^t$ is the local map shape in the frame t, and $\mathbf{T}_{cw}^t$ the camera pose. Black points belong to the global map. Some of them are embedded in the template. Current matched points in red.

the focal lengths and the principal points from the camera calibration. The set of observation in the image $\mathcal{I}^t$ is the keypoints x^t matched with a map point of $\mathcal{X}^t$. The map point $\mathbf{X}_j^t$ is projected in the normalized retina as $(\hat{x}_j^t, \hat{y}_j^t)$ where $\hat{x}_j^t = \frac{x_j^t - C_x}{f_x}$, $\hat{y}_j^t = \frac{y_j^t - C_y}{f_y}$ and $(x_j^t y_j^t)^\top = \pi(\mathbf{T}_{tw}, \mathbf{X}_j^t)$.

C. Camera Pose and Template Deformation

In SLAM sequences, the camera usually images a zone smaller than the template. For efficiency and scalability, we only optimize the observed zone of the template and its closest vicinity. We refer to this part of the template as the local zone $\mathcal{L}_k^t \subseteq \mathcal{T}_k^t$. Fig. 2 shows all the components of the deformable tracking: the template $\mathcal{T}_k^t$, the local zone $\mathcal{L}_k^t$ and the camera pose $\mathbf{T}_{tw}$.

To estimate the deformed $\mathcal{L}_k^t$ and $\mathbf{T}_{tw}$, we jointly minimize the reprojection error $\varphi_d(\mathcal{I}^t, \mathbf{T}_{cw}, \mathcal{L}_k^t)$ in the image I^t and the deformation energy $\varphi_e(\mathcal{L}_k^t, \mathcal{T}_k)$ of the template $\mathcal{T}_k$:

$$\underset{\mathcal{L}_k^t, \mathbf{T}_{tw}}{\arg\min} \quad \varphi_d(\mathcal{I}^t, \mathbf{T}_{tw}, \mathcal{L}_k^t) + \varphi_e(\mathcal{L}_k^t, \mathcal{T}_k). \tag{3}$$

We solve (3) using the Levenberg–Marquardt optimization method. The initial guess for $(\mathcal{L}_k^t, \mathbf{T}_{tw})$ is the solution of the previous frame $(\mathcal{L}_k^{t-1}, \mathbf{T}_{t-1w})$. We fix the pose boundary nodes of $\mathcal{L}_k^t$ during the optimization to constraint the gauge freedoms of the camera pose $\mathbf{T}_{tw}$.

The reprojection error $\varphi_d(\mathcal{I}^t, \mathbf{T}_{tw}, \mathcal{L}_k^t)$ for the set of keypoints x^t in image $\mathcal{I}^t$ is defined as

$$\varphi_d(\mathcal{I}^t, \mathbf{T}_{tw}, \mathcal{L}_k^t) = \sum_{j \in x^t} \rho\left(\left\|\pi(\mathbf{X}_j^t, \mathbf{T}_{tw}) - \mathbf{x}_j^t\right\|\right). \tag{4}$$

The reprojection error is robust against outliers as it is weighted with a Huber robust kernel $\rho(.)$.

We define a deformation energy $\varphi_e(\mathcal{L}_k^t, \mathcal{T}_k)$ with respect to $\mathcal{T}_k$ as a combination of a stretching energy $\varphi_s(\mathcal{L}_k^t, \mathcal{T}_k)$, a bending energy $\varphi_b(\mathcal{L}_k^t, \mathcal{T}_k)$, and a reference regularizer $\varphi_r(\mathcal{L}_k^t, \mathcal{T}_k)$

$$\begin{aligned}\varphi_e(\mathcal{L}_k^t, \mathcal{T}_k) = \; & \lambda_s \varphi_s(\mathcal{L}_k^t, \mathcal{T}_k) + \lambda_b \varphi_b(\mathcal{L}_k^t, \mathcal{T}_k) \\ & + \lambda_r \varphi_r(\mathcal{L}_k^t, \mathcal{T}_k).\end{aligned} \tag{5}$$

We use λ_s, λ_b, and λ_r to weight the influence of each term.

The stretching energy $\varphi_s(\mathcal{L}_k^t, \mathcal{T}_k)$ measures the difference in the length l_e^t of each edge e in the local zone $\mathcal{L}_k^t$ in frame t with respect to its length l_e^k in the shape-at-rest of $\mathcal{T}_k$

$$\varphi_s(\mathcal{L}_t^k, \mathcal{T}_k) = \sum_{e \in \mathcal{L}_k^t} \left(\frac{l_e^t - l_e^k}{l_e^k}\right)^2. \tag{6}$$

The bending energy $\varphi_b(\mathcal{L}_k^t, \mathcal{T}_k)$ measures the changes in mean curvature δ_n^t in each node n with respect to the estimated δ_n^k in the shape-at-rest of $\mathcal{T}_k$. We estimate the mean curvature through the discrete Laplacian operator [36]. We make the bending term dimensionless by dividing it by the mean distance l_e^k of the edges connected with the node $\mathcal{E}_n^k$

$$\varphi_b(\mathcal{L}_k^t, \mathcal{T}_k) = \sum_{n \in \mathcal{L}_k^t} \sum_{e \in \mathcal{E}_n^k} \left(\frac{\delta_n^t - \delta_n^k}{l_e^k}\right)^2. \tag{7}$$

Optimization considering the terms $\varphi_d(\cdot)$, $\varphi_b(\cdot)$, and $\varphi_s(\cdot)$ allows to recover the relative pose of the camera with respect to the template, but the absolute camera pose is not observable. Thanks to the fixation of the $\mathcal{L}_k^t$ boundary nodes pose, the absolute camera pose becomes observable. However, the camera pose sometimes is only weakly observable depending on the boundary nodes geometrical distribution and cardinality. If the template is completely observed by the camera, then there are no boundary points to be fixed and the camera pose becomes fully nonobservable.

We add another regularizer, $\varphi_r(\mathcal{L}_k^t, \mathcal{T}_k)$, that we call reference regularizer to keep the template as close as possible to its initial position in its reference keyframe, to alleviate the camera pose weak observability. It is given by

$$\varphi_r(\mathcal{L}_k^t, \mathcal{T}_k) = \sum_{n \in \mathcal{L}_k^t} \left\|\mathbf{V}_n^t - \mathbf{V}_n^k\right\|. \tag{8}$$

Optimization (3) also needs the derivatives of the regularizers (6)–(8); they are detailed in Appendix A.

D. Data Association

To match the keypoints in the current frame with the map points, we apply an active matching strategy as proposed in [37]. First, the ORB keypoints are detected in the current frame. Next, the camera pose is predicted with a camera motion model as a function of the past camera poses. Then, we use the last estimated shape of template and the barycentric coordinates to predict where the map points will be imaged. Around the map point prediction, we define a search region. We match the map point with the keypoint with the most similar ORB descriptor inside its search region. The similarity is estimated as the Hamming distance between the ORB descriptors, and the match is accepted only if it is below a distance threshold. The ORB descriptor of the map point is taken from the keypoint of the keyframe where it was initialized.

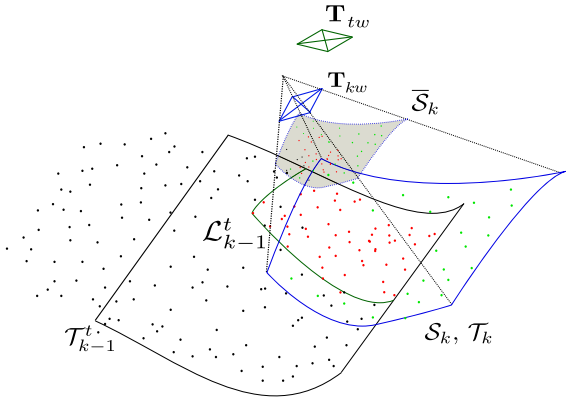


Fig. 3. Extension of the map in the deformable mapping. Local area $\mathcal{L}^t_{k-1}$ in green. Matched points in red. In blue, the up-to-scale surface estimated by NRSfM, $\overline{\mathcal{S}}_k$ (dotted line), and template $\mathcal{T}_k$ computed from the scaled surface $\mathcal{S}_k$ of the reference keyframe k.

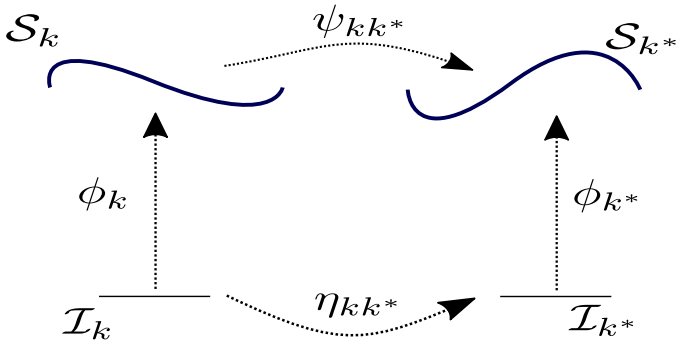


Fig. 4. Relation between an anchor keyframe k and one of its covisibles k^*. ϕ_k and ϕ_{k^*} are embeddings of the two keyframe surfaces k and k^*. η_{kk^*} is the warp between k and k^*. ψ_{kk^*} is the deformation field between the surfaces S_k and S_{k^*}.

E. New Keyframe Selection

We select a new keyframe as soon as the mapping thread finishes its processing. If the new keyframe covers a new map region, it becomes an anchor keyframe and the reference keyframe, and a new template is created. Otherwise, the new keyframe is a regular keyframe, and its most covisible anchor keyframe is selected as the reference keyframe, and its template is refined.

V. DEFORMATION MAPPING

Deformation mapping recovers the observed map as a surface S_k for the reference keyframe k. This surface contains the map points observed in the keyframe during the tracking. With the new keyframe, we refine the map points and create new ones. S_k defines the shape-at-rest of the template $\mathcal{T}_k$ for the deformation tracking for the next frames, as shown in Fig. 3.

Deformation mapping is performed as follows. First, we compute the warps η_{kk^*} between the anchor keyframes k and the new keyframe k^*. At this stage, the considered anchor keyframes are those where one of the currently observed map points were intialized. Second, we estimate an up-to-scale surface $\overline{\mathcal{S}}_k$ by processing the covisible keyframes with the new keyframe by means of NRSfM. Third, we align $\overline{\mathcal{S}}_k$ with the previous map to recover the scale and the scaled surface $\mathcal{S}_k$. Finally, with this new surface, we create the new template by computing a triangular mesh and embedding the map points in its facets.

A. NRSfM

In isometric NRSfM, the surface deformation is modeled locally for each point under the assumption of isometry and infinitesimal planarity. Assuming infinitesimal planarity, any surface is approximated as a plane at an infinitesimal level, while maintaining its curvature at the global level. Isometric NRSfM can handle both rigid and nonrigid scenes. Since we use a local method, it can handle missing data and occlusions inherently. We build on the isometric NRSfM proposed in [6]. For the sake of completeness, we summarize the formulation.

ϕ_k is the embedding of the scene surface $\mathcal{S}_k$; it is parametrized using the retina normalized coordinates of the image $\mathcal{I}_k$

$$\phi_k : \mathbb{R}^2 \mapsto \mathbb{R}^3$$
$$\phi_k(\hat{x},\hat{y}) = \left[\frac{\hat{x}}{\beta(\hat{x},\hat{y})} \quad \frac{\hat{y}}{\beta(\hat{x},\hat{y})} \quad \frac{1}{\beta(\hat{x},\hat{y})}\right]^\top \tag{9}$$

where $\beta_k(\hat{x},\hat{y})$ is the inverse depth of each point. The normal $\vec{\mathbf{n}}_j(\hat{x},\hat{y})$ of the surface expressed with respect to this parametrization is given as

$$\vec{\mathbf{n}}_j(\hat{x},\hat{y}) \propto \begin{pmatrix} K_{\hat{x}} \\ K_{\hat{y}} \\ 1 - \hat{x}K_{\hat{x}} - \hat{y}K_{\hat{y}} \end{pmatrix} \tag{10}$$

where $K_{\hat{x}} = \frac{\beta_k(\hat{x},\hat{y})_{\hat{x}}}{\beta_k(\hat{x},\hat{y})}$ and $K_{\hat{y}} = \frac{\beta_k(\hat{x},\hat{y})_{\hat{y}}}{\beta_k(\hat{x},\hat{y})}$, and the subindexes $\hat{x}$ and $\hat{y}$ denote the partial derivatives.

NRSfM exploits the relationship between the metric tensor, $g_k(\hat{x},\hat{y})$, and the Christoffel symbols, $\Gamma^{\hat{x}}_k(\hat{x},\hat{y})$ and $\Gamma^{\hat{y}}_k(\hat{x},\hat{y})$, of the surface of the keyframe $\mathcal{S}_k$ and those of its covisible keyframes $\mathcal{S}_{k^*}$. Assuming infinitesimal planarity and isometry, $\Gamma^{\hat{x}}_k(\hat{x},\hat{y})$ and $\Gamma^{\hat{y}}_k(\hat{x},\hat{y})$ only depend on $K_{\hat{x}}$ and $K_{\hat{y}}$ for each point in every keyframe image. The warp η_{kk^*} between the keyframes k and k^* represents the transformation from image $\mathcal{I}_k$ to image $\mathcal{I}_{k^*}$. Fig. 4 shows the different elements of the two view relation, the warp η_{kk^*}, the surface embeddings for each keyframe ϕ_k and ϕ_{k^*}, and the isometric deformation ψ_{kk^*} between the surfaces $\mathcal{S}_k$ and $\mathcal{S}_{k^*}$. Due to the infinitesimal planarity and isometry assumptions, the metric tensor and the Christoffel symbols in two different surfaces k and k^* are related through the warp between these keyframes η_{kk^*} as

$$g_k(\hat{x},\hat{y}) = J^\top_{\eta_{kk^*}} g_{k^*}(\hat{x}^*,\hat{y}^*) J_{\eta_{kk^*}} \tag{11}$$

$$\Gamma^q_k(\hat{x},\hat{y}) = \sum_h \frac{\partial \hat{x}_h}{\partial \hat{x}^*_h}\left(J^\top_{\eta_{kk^*}} \Gamma^h_k(\hat{x}^*,\hat{y}^*) J_{\eta_{kk^*}} + H^h_{\eta_{kk^*}}\right) \tag{12}$$

where $J_{\eta_{kk^*}}$ and $H^q_{\eta_{kk^*}}$ are the Jacobian and the Hessian for the variable $q = \{\hat{x},\hat{y}\}$ of the warp η_{kk^*} respectively. Equations (11) and (12) can be transformed in two cubic polynomial equations

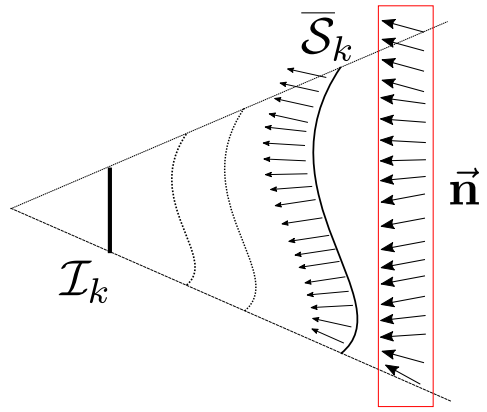


Fig. 5. $\overline{\mathcal{S}}_k$ is the estimated up-to-scale surface. $\vec{\mathbf{n}}$ are the set of normals. Two examples of surfaces at a different scale but having the same normals are displayed in dotted lines.

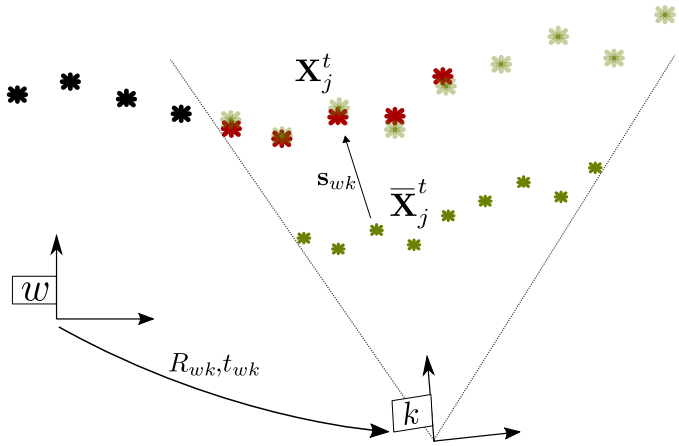


Fig. 6. Sim(3) alignment. We align the the map points $\overline{\mathbf{X}}_j^k \in \overline{\mathcal{S}}_k$ of the up-to-scale estimation with the pose of the map points $\mathbf{X}_j^k \in \mathcal{T}_{k-1}^k$ estimated for the frame k deforming the previous template $k-1$.

$P(K_{\hat{x}}^k, K_{\hat{y}}^k)$ and $Q(K_{\hat{x}}^k, K_{\hat{y}}^k)$ for each point correspondence

$$P(K_{\hat{x}}^k, K_{\hat{y}}^k) = \sum_{u,v\in[0,3]} p_{uv}(K_{\hat{x}}^k)^u (K_{\hat{y}}^k)^v = 0 \qquad (13)$$

$$Q(K_{\hat{x}}^k, K_{\hat{y}}^k) = \sum_{u,v\in[0,3]} q_{uv}(K_{\hat{x}}^k)^u (K_{\hat{y}}^k)^v = 0 \qquad (14)$$

where the coefficients p_{uv} and q_{uv} depend only on the normalized coordinates of the points and the derivatives of first- and second-order derivatives of the warp η_{kk^*}. We refer to [6] for further details in the coefficients p_{uv} and q_{uv}.

B. Incremental Surface Normals Refinement

If a point is matched in two or more keyframes, we can calculate its normal in its anchor keyframe k, defined by $K_{\hat{x}}^k$ and $K_{\hat{y}}^k$, by means of nonlinear optimization

$$\underset{K_{\hat{x}}^k, K_{\hat{y}}^k}{\arg\min} \left(P\left(K_{\hat{x}}^k, K_{\hat{y}}^k\right)\right)^2 + \left(Q\left(K_{\hat{x}}^k, K_{\hat{y}}^k\right)\right)^2. \qquad (15)$$

In contrast to [6], optimization (15) is incrementally computed. We initialize it with its last estimate achieving a fast convergence. Once the normals are refined in their anchor keyframe, we transfer the normals to the new reference keyframe with (12). We recover the up-to-scale $\overline{\mathcal{S}}_k$ from the set of estimated normals $\vec{\mathbf{n}}$ using Shape-from-Normals (SfN) [4]. The surface $\overline{\mathcal{S}}_k$ is regressed with a bicubic b-spline parametrized by its control nodes depth. The control nodes are defined by a regular mesh in the image $\mathcal{I}_k$. We fit the depth of the nodes to obtain a surface orthogonal to the estimated normals with a regularizer in terms of bending energy (Fig. 5).

C. Surface Alignment

The new estimated surface $\overline{\mathcal{S}}_k$ is up-to-scale. We need to recover the solution with a coherent scale s_{wk} with respect to the already estimated map. This means that the scale-corrected shape-at-rest $\mathcal{S}_k$ must have a scale coherent with the deformed template $\mathcal{T}_{k-1}^k$ estimated by the tracking when the keyframe was inserted.

We align these surfaces map points through a transformation which belongs the group of similarity of 3-space Sim(3), by means of nonlinear optimization:

$$\underset{\mathbf{R}_{wk}, \mathbf{t}_{wk}, \mathbf{s}_{wk}}{\arg\min} \sum_{j\in\mathbf{X}^k} \left\| \mathbf{s}_{wk}\mathbf{R}_{wk}\overline{\mathbf{X}}_j^k + \mathbf{t}_{wk} - \mathbf{X}_j^k \right\|^2, \qquad (16)$$

where $\mathbf{R}_{wk}$, $\mathbf{t}_{wk}$, and $\mathbf{s}_{wk}$ are the rotation translation and scale defining the Sim(3) transformation (Fig. 6).

To build our new template $\mathcal{T}_k$, we finally create a triangular mesh from the scale-corrected surface $\mathcal{S}_k$ by means of regular triangular mesh in the image. The new map points 3-D pose is computed from the matched keypoints by constraining them to be in the estimated surface $\mathcal{S}_k$. Then, we embed the re-observed map points and the new map points by projecting them into their corresponding template facet. With this embedding, we calculate the barycentric coordinates of the map points which will be used by the tracking.

D. Template Substitution

Once the surface S_k is computed, the keyframe k is set as the reference keyframe and the current template $\mathcal{T}_{k-1}$ is substituted by $\mathcal{T}_k$ computed from S_k. The shape observed in the current frame t differs from the shape of the new template $\mathcal{T}_k$. This yields to failures in the data association stage, which assumes small deformations, if we substitute the template directly by $\mathcal{T}_k$. Therefore, we transfer the matches from $\mathcal{T}_{k-1}^t$ to $\mathcal{T}_k$ and compute the current shape $\mathcal{T}_k^t$ using optimization (3).

E. Warp Estimation and Nonrigid Guided Matching

The input of NRSfM is the set of warps η_{kk^*} between an anchor keyframe k and their covisible keyframes k^*. The image warp η_{kk^*} is a function that transforms a point in the anchor

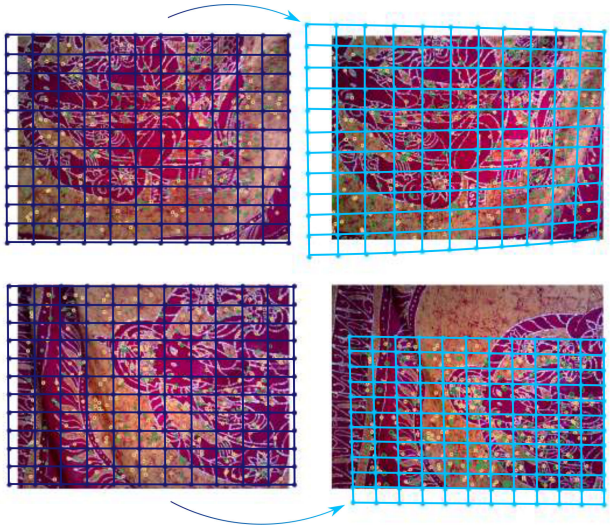


Fig. 7. Two examples of warp estimation. Warp estimation between the keyframe k (left) and k^* (right). The warp between k and k^* is plotted in blue. Yellow points are the initially matched map points and green points are the matches added by guided matching stage using the warp.

Fig. 8. Two configurations of Mandala dataset: rigid planar (mandala0), and hanged in the rest of the sequences.

keyframe into the corresponding point in its covisible k^*:

$$\eta_{kk^*} : [\hat{x}, \hat{y}] \in \mathbb{R}^2 \mapsto \ [\hat{x}^*, \hat{y}^*] \in \mathbb{R}^2.$$

We use a particular family of warps called Schwarps [38], because, as discussed in [6], the formulation of the 2-D Schwarzian equation regularizers is equivalent to the infinitesimal planarity of the NRSfM. See Fig. 7 for two examples of warp between keyframes.

First, we estimate an initial warp between the anchor keyframe k and its covisible keyframe k^* with the matches given by the deformation tracking. Then, we use the initial warp to perform a guided matching stage between the keypoints in keyframes k and k^*. We accept as a match the keypoint inside a search region with the smallest Hamming distance for the ORB descriptor. We apply a threshold on the ORB similarity to definitively accept a match. Once that we have the new matches, we incorporate them to the initial ones and estimate the final warp. See Fig. 7 for two examples of warp between keyframes.

F. SLAM Initialization

At initialization, we need to have a template available for the scene surface. We compute it from the first frame of the sequence, assuming its surface $\mathcal{S}_1$, and hence its template $\mathcal{T}_1$ is a plane parallel to perpendicular to the camera optical axis.

With the second keyframe inserted, the mapping thread starts to compute a new template that replaces the initial one. The accuracy of the first computed templates strongly depends on how many keyframes are fed in the NRSfM and on how large is the parallax they render.

According to the experiments, our algorithm can track from an inaccurate template with a high-quality data association between keyframes, yielding long tracks and a low false positive rate. As a result, as more keyframes rendering high parallax are created, the estimated template eventually converges to the actual scene shape.

VI. Implementation Details

The method is implemented in C++ and runs entirely on the CPU. We have used the OpenCV library [39] for base computer vision functions. For the SfT optimization and the LS $\mathrm{Sim}(3)$ registration, we have used the g2o library [40] and its implementation of Levenberg–Marquardt. For the Schwarps optimization, the normal estimation, and the SfN, we have used the Ceres library [41]. The runtime depends on the resolution of the mesh used as template. For a mesh of 10×10 nodes, the runtime is approximately 50 ms for the deformable tracking thread and approximately 400 ms for the deformable mapping in a machine with an i7-4700HQ CPU and with 7.7 Gb RAM. The code will is available as a public git repository.[1]

VII. Experiments

We tested DefSLAM in two datasets. The first dataset is the Mandala dataset which we create to evaluate deformable monocular SLAM in a laboratory-controlled situation. The second is a selection of sequences from the medical Hamlyn dataset ([13], [42]), which comprises a phantom heart, and *in vivo* sequences including exploratory trajectories. The sequences in both datasets have ground truth depth for each frame, either from stereo or from CT.

We focus on two per frame metrics: the 3-D rms error of the in-frustum map points and the fraction of matched map points. The rms error is computed after a scale alignment for each frame of the sequence, which features the geometrical accuracy. The fraction of map points matched is the quotient between the map points effectively matched in the current frame, and the number of map points in-frustum of the current frame, i.e., maximum number of map points that ideally can be matched. A low fraction signals a poor map that can only represent partially the scene imaged in the current frame.

In addition, we carried out an ablation analysis of the mapping and the tracking. In the mapping, we focused in NRSfM stages of the normals estimation. In the tracking, we evaluate the

[1][Online] Available: https://github.com/UZ-SLAMLab/DefSLAM

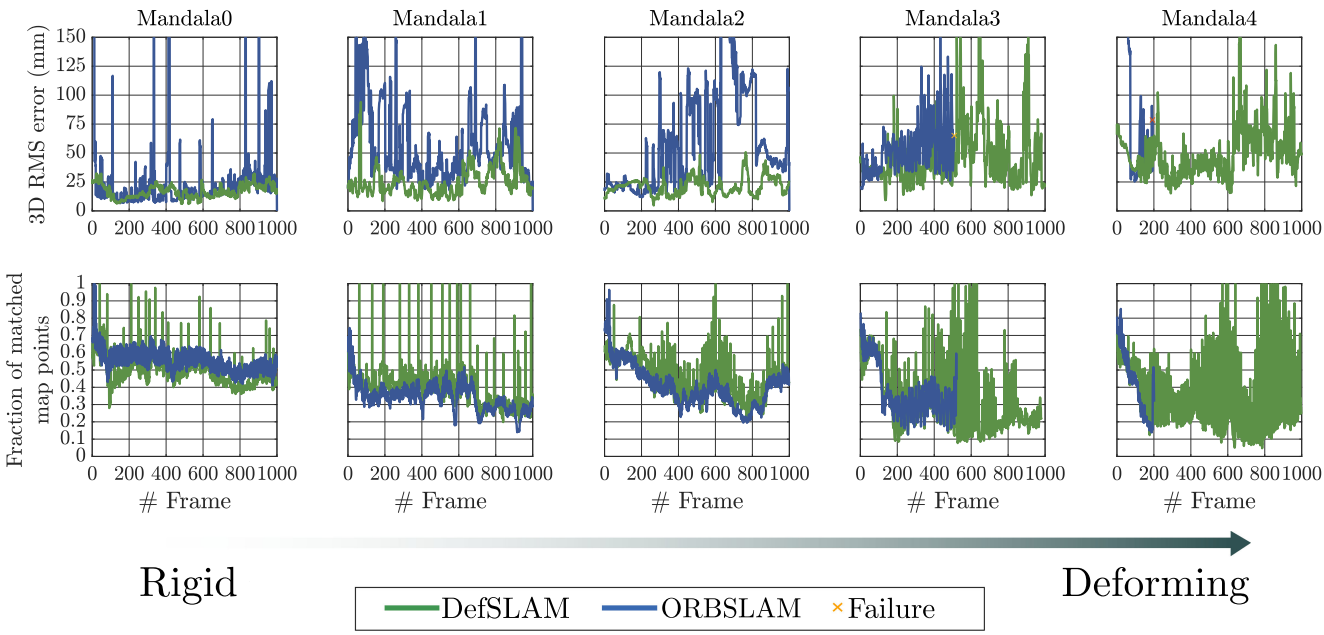


Fig. 9. Overall quality for Mandala dataset sequences. From left to right, the scenario contains more deformation. Top: 3-D rms error (mm) per frame (the smaller, the better). Bottom: Fraction of matched map points (the higher, the better).

performance of the deforming template when compared with a rigid one. We also analyzed the sensitivity of the system to the tuning of the regularizers' weights in the tracking optimization (5).

Currently, there is no other monocular SLAM for deformable environments to compare with. Thus, we select a rigid monocular SLAM method, ORBSLAM [3], as one of the closest for comparison. We had to retune several stages of ORBSLAM to process deforming sequences. 1) We relaxed the thresholds for matching and outlier rejection to retain matches despite the deformation. 2) We initialized it with the first frame ground truth map, to avoid the dramatical failure of the monocular intialization. 3) We decreased the rate of new keyframe creation up to one keyframe out of three frames, to adapt the map to the scene deformations. On the other hand, we compare with MISSLAM [14] in the Hamlyn phantom heart dataset, as the closest in the medical arena, despite MISSLAM being a stereo instead of monocular.

For the sake of repeatability, DefSLAM was run sequentialized in single-thread, inserting one new keyframe every ten frames. All the reported results are the median of five executions in each sequence.

A. *Mandala Dataset*

We introduce the *Mandala dataset* to evaluate the map quality of deformable monocular SLAM systems in a controlled environment. It is composed of five sequences (640 × 480 pix. at 30 fps) with exploratory trajectories observing a textured kerchief deforming near-isometrically. We increased the hardness of deformation progressively by reducing the period of the waves generated on the kerchief and increasing their amplitude from the shape-at-rest. Fig. 8 shows the two configurations: planar and hanged.

In the sequence mandala0, the kerchief remains rigid on the floor. In mandala1, the deformation had an amplitude of 15 cm and a period of 2 s. In mandala2, the amplitude is 10 cm and the period 1 s. In mandala3, the amplitude is 25 cm and the kerchief oscillates with a period of 2 s. In mandala4, the amplitude is 30 cm, and its period is halved to 1 s.

1) Overall Quality Experiment: We analyze the overall quality of the estimated map. Fig. 9 shows the final results along the five sequences for DefSLAM in green, and ORBSLAM in blue.

In rigid mandala0, DefSLAM obtains a similar 3-D rms error to ORBSLAM. Concerning the fraction of matched map points, both DefSLAM and ORBSLAM got a high percentage, which means that the map points are highly reused due to the rigidity of the scene.

In mandala1 and mandala2, the kerchief has low frequency and amplitude deformation. DefSLAM obtains a similar 3-D rms error to the one obtained in mandala0 for both sequences, being able to recover the deformation of the kerchief. ORBSLAM could process the entire sequences, but its 3-D rms error was highly penalized by the deformation, triplicating the error obtained in the mandala0 sequence, and the rms error of DefSLAM. DefSLAM could recover more accurately the deformation of the scene observed during the sequence both in terms of rms error and in fraction of matched map points per frame.

In the mandala3 and mandala4 sequences, the conditions are more extreme. ORBSLAM could not process any of these sequences entirely. In this sequences, the fast deformation yields difficulties for DefSLAM which experiments some delay to converge the correct shape. This provoked some peaks in the rms error. In any case, the error average was around the 4 cm during both sequences. In Fig. 10, we can observe the quality

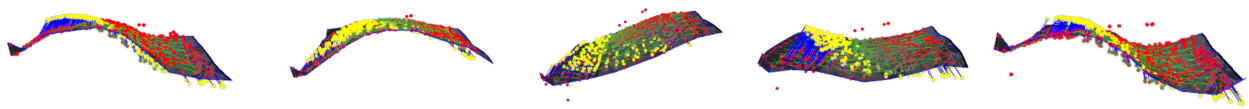


Fig. 10. Recovering local deformations in the mandala3 sequence. 3-D map points in red, 3-D point in yellow is the ground truth, and blue lines are the difference. DefSLAM can perceive and reconstruct the deforming scene.

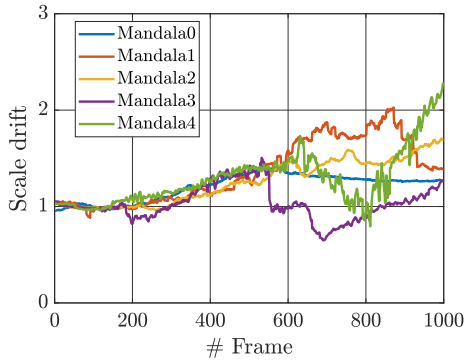


Fig. 11. Scale drift along the Mandala sequences. It increases more with more challenging. It is reduced in case of reobservation.

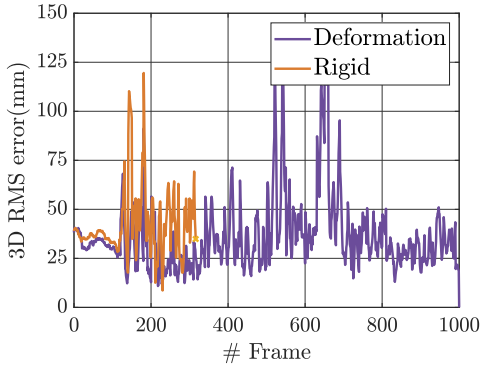


Fig. 12. Rigid tracking vs. deformation tracking surface error as 3-D rms scene reconstruction error per frame in mm.

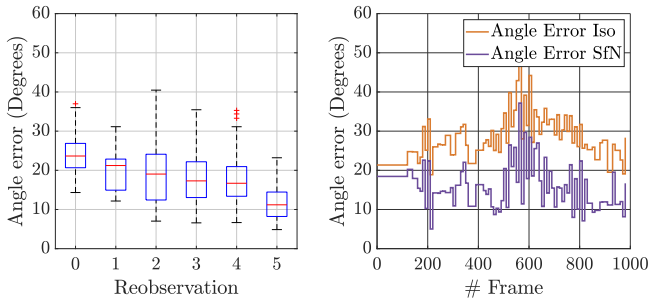


Fig. 13. (Left) Box-and-whisker plot for the normals angle error in a keyframe after SfN, improvement as a function of the keyframe resobservations. (Right) Per keyframe RMSE angle error for the normal orientation after NRSfM and after SfN.

of the reconstruction of the local deformations in the sequence mandala3. The fraction of matched map points for DefSLAM was also smaller. Supplementary material includes a video with fragments of the mandala dataset quality results.

2) Scale Drift Analysis: The previous section RMSE focuses on the up-to-scale shape accuracy. Fig. 11 shows the scale drift along the different sequences. The main source of scale drift is the alignment (Section V-C), where to estimate the scaled template, we align the reference up-to-scale template with the previous reference scaled template. This makes the scale accumulate the misalignment between the new and the old templates. The scale drift is around a 10% in the mandala 0 and increases to higher values to peak the deformation becomes more challenging. Eventually, the scale drift can be reduced due to reobservations of the map during the sequence.

3) Sensitivity Analysis: All the experiments reported, both in the Mandala dataset and Hamlyn, were run with $\lambda_s = 16\,000$, $\lambda_b = 300$ and $\lambda_r = 0.02$ as standard tuning.

To better understand the role of the weights, we varied their values to study their effect in the final 3-D rms error and scale drift in the challenging mandala3. We run the entire sequence and evaluated the rms error at the end of the sequence from frames # 800 to # 1000. The error is not severely affected, remaining between 20 and 40 m, for a range of values from $\lambda_s = [1600, 10\,0000]$, $\lambda_b = [100, 1000]$, and $\lambda_r = [0, 0.1]$. By decreasing the λ_s and λ_b values, the system becomes unconstrained and fails in processing the entire sequence. By increasing λ_s and λ_b, the system assumes rigidity, thus causing another failing scenario. Fig. 12 shows the extreme case of a perfectly rigid and fixed template compared with our standard tuning. It can be seen how a rigid template for tracking fails to survive strong scene deformations. This case corresponds to high values for the three coefficients λ_s, λ_b, and λ_r.

The reference regularizer has proven critical to reduce the scale drift specially in the Hamlyn SeqHeart sequence where the camera is imaging constantly the same zone and observing the entire template with few boundary point constraints (Section VII-B), from 36% for λ_r to 2% for $\lambda_r = 0.02$.

4) Deformation Mapping Normal Estimation Accuracy: We analyze the quality of the deformation mapping for sequence mandala3 focusing in the angle error between the estimated normal and the ground truth normal, in the two stages of the normal estimation, the initial NRSfM and the subsequent SfN (Section V-B).

Fig. 13 shows the rms angle error of the shape estimated by the NRSfM versus the error after the SfN stage. SfN consistently reduces the error through the entire sequence improving the normals. Averaging the error for all the keyframes in the sequence, the SfN achieves a 15-deg RMSE versus the 22 deg of the NRSfM.

The output of the NRSfM is the set of surface normals for each map point in the reference keyframe. The normal of a map

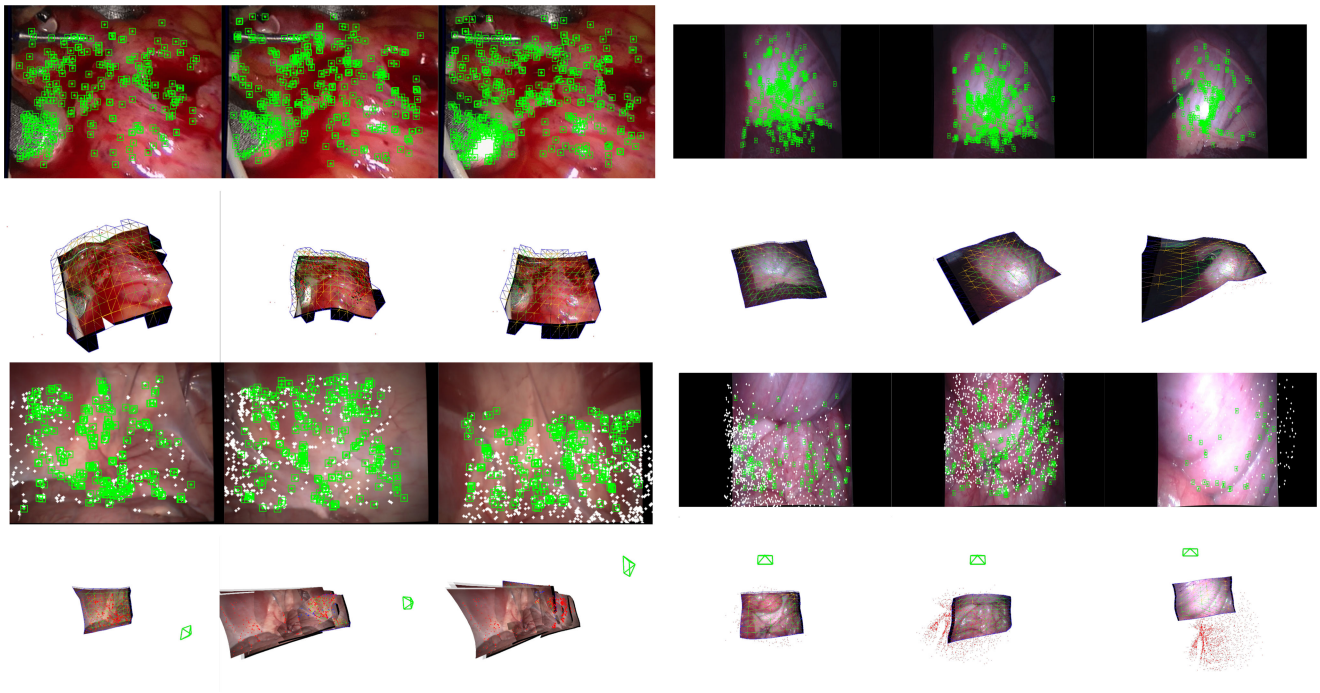


Fig. 14. DefSLAM in *in vivo* Hamlyn dataset sequences. Three typical 2-D images an the corresponding 3-D maps. (Top left) Heart sequence. (Top right) Organs Sequence. (Bottom left) Abdominal sequence. (Bottom right) Exploration sequence.

point is re-estimated after each reobservation of that point in a new keyframe. Fig. 13 shows the evolution of the rms angle error for the normals in a keyframe along five reobservations after its creation. We can see how the median error goes from 23 deg at initialization down to 12 deg after the fifth reobservation.

B. Hamlyn Dataset

Our last experiments test DefSLAM in intracorporeal in six sequences from the *Hamlyn dataset* [13], [42] to evaluate our algorithm in medical images. The first two sequences are recorded with an *ex vivo* phantom heart [43] synchronized with a CT scanner to register ground truth. In addition, we processed four *in vivo* laparoscopic sequences (see Fig. 14): 1) SeqAbdomen is an exploration of the abdominal wall where the scene remains almost rigid (Fig. 14, bottom left). 2) SeqExploration performs an exploration around the exterior of the bowel with low texture. It has a small deformation at the beginning (Fig. 14, bottom right). 3) SeqHeart [42] is a nonrigid beating heart observed by a fixed camera. 4) SeqOrgans is an abdominal exploration and deformation of the scene due to tool interfering (Fig. 14, top right).

The closest SLAM system to ours reporting accuracy with respect to an external sensor in medical sequences is MISSLAM [14]. We evaluate our system in the same sequences, i.e., the *ex vivo* phantom heart sequences. Despite the lack of camera motion, the scenes have enough deformation for DefSLAM to reconstruct them. We report a mean accuracy of 3 and 4 mm in the sequence phantom5 and phantom7, respectively. The average accuracy MISSLAM as reported by the authors in [14] is 0.28 and 0.35 mm. Concerning the execution time, we report a similar runtime per frame, but DefSLAM runs in CPU unlike MISSLAM that uses GPU. It has to be noted that they use stereo input in contrast with DefSLAM which is a purely monocular method.

Fig. 15 reports the median of five executions rms error during the four *in vivo* Hamlyn sequences and Fig. 16 shows its corresponding scale drift. As it happened with the Mandala dataset (Section VII-A2), the scale drift got slightly increased for the more challenging sequence. In the sequence where the camera is in the same zone, there is no scale drift.

DefSLAM is able to process SeqAbdomen and SeqExploration entirely with a mean 3-D rms error of 17 and 10 mm, respectively. In these scenes, the camera explores but it comes back to the same zone. DefSLAM was able to reobserve part of the map already built and thus reduced the scale drift. ORBSLAM performed poorly in this sequences and could not process them entirely.

In SeqHeart, the camera is practically static, but DefSLAM was able to initialize with the monocular strategy proposed even with a short parallax. The 3-D rms error was approximately 3 mm, equal to the *ex vivo* phantom result with a much better ground truth. ORBSLAM initialized with the ground truth was able to process the entire sequence with an error of 5 mm.

Finally, in the sequence SeqOrgans, DefSLAM shows its ability to perform the reconstruction of a deformable scene in exploratory sequence with an accuracy of 8 mm. It survives to the tool clutter that covers almost entirely the image, correcting the scale drift. In the end of the sequence, the tool deforms the organ imaged and DefSLAM was able to recover the deformation of the scene with the same error than in the rest of the sequence.

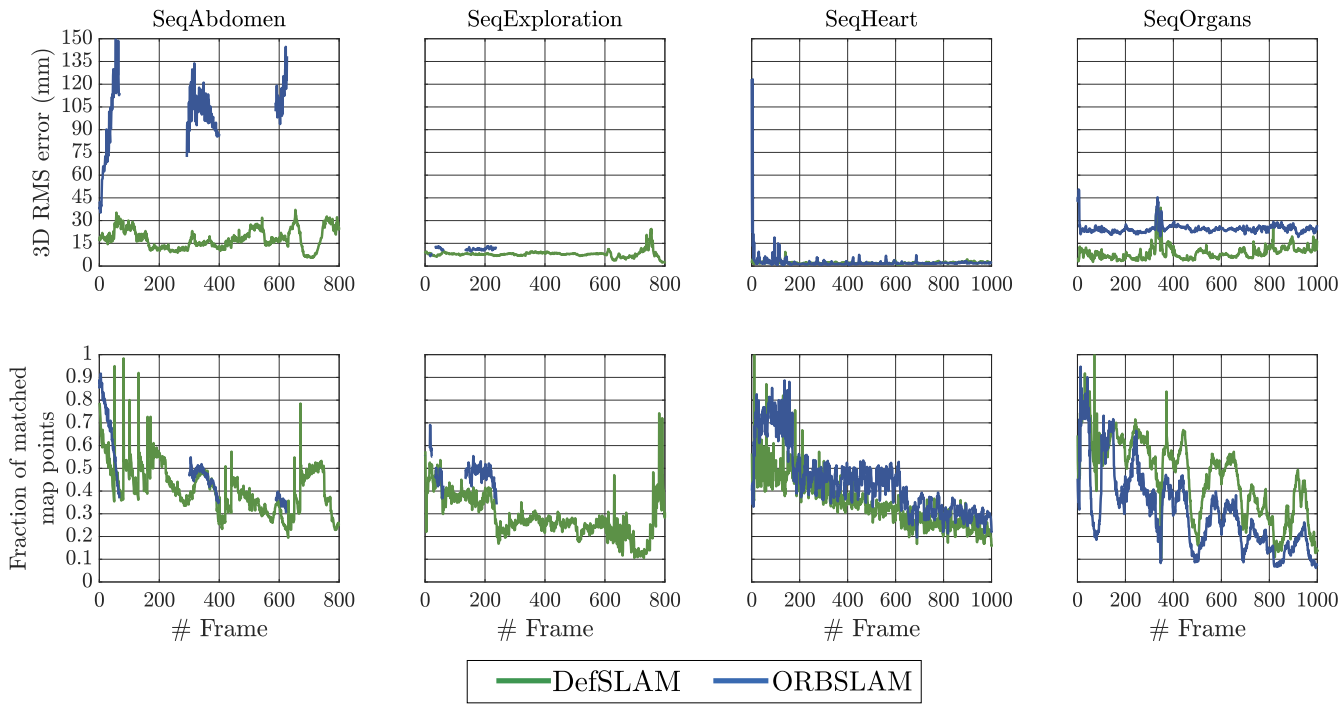


Fig. 15. Processing Hamlyn sequences. Green DefSLAM, blue ORBSLAM. From left to right: Heart, organs, abdomen and exploration sequences. Per frame rms scene reconstruction error in mm after a per frame scale alignment with the stereo ground truth.

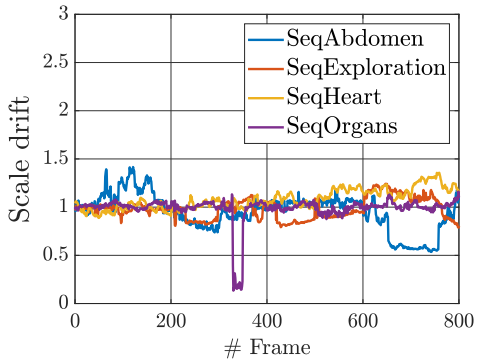


Fig. 16. Scale drift along the Hamlyn dataset sequences.

Fig. 14 shows the overall quality of the 3-D reconstruction of the medical sequences. Supplementary material includes the video with the results in all the sequences.

VIII. Conclusion

In this article, we formulated DefSLAM, the first deformable SLAM able to process monocular sequences. We proposed to split the computation of DefSLAM into two parallel threads. The deformation tracking thread was devoted to estimating the camera pose and the deformation of the scene, which was based on SfT. SfT needs a prior of the geometry of the scene encoded in the template. When exploring new zones, our method estimated new templates to cover new areas. Our second thread, the deformation mapping, was devoted to periodically re-estimating the template to better adapt it to the currently observed scene. Both SfT and NRSfM model the cameras as perspective, hence the system was able to handle close-ups typical in scene exploration where perspective effects were prevalent.

Our experiments confirm that the proposed method was able to handle real exploratory trajectories of a deforming scene. Since direct comparison with other systems was not possible, we focused the comparison with the rigid monocular ORBSLAM after its retuning to handle nonrigid scenes. This comparison proved that DefSLAM was able to robustly initialize from monocular sequences, continuously adapt the map to the scene deformation, and produce accurate scene estimates.

We also showed in preliminary experiments that the system was able to handle medical endoscopy images. The next step will be its adaptation for medical imagery to handle all kinds of challenges not taken into account in the present work, i.e., uneven illumination, poor visual texture, and nonisometric deformations or ultra close-up shots exploring the endoluminal cavities.

Another future work is to develop a full-fledged mapping system including multiple maps, relocalization, loop closure, or long-term place recognition to achieve robust performance for extended periods of time or multiple moving and deforming bodies.

Appendix

A. Derivatives of Regularizers

We show the Jacobian terms of the regularizers to prove that they do not have singularities:

1) Stretching:

The streching error $e_s(\mathcal{L}_t^k, \mathcal{T}_k)_e$ for the edge e is

$$e_s(\mathcal{L}_t^k, \mathcal{T}_k)_e = \left(\frac{l_e^t - l_e^k}{l_e^k} \right) \tag{17}$$

being

$$l_e^t = \|(\boldsymbol{V}_{e^1}^t - \boldsymbol{V}_{e^2}^t)\|_2 \tag{18}$$

where $\boldsymbol{V}_{e^1}^t$ and $\boldsymbol{V}_{e^2}^t$ are the two nodes of the edge e in instant t. Its derivative is

$$\frac{\partial e_s(\mathcal{L}_t^k, \mathcal{T}_k)_e}{\partial \boldsymbol{V}_{e^i}^t} = \frac{(\boldsymbol{V}_{e^1}^t - \boldsymbol{V}_{e^2}^t)}{l_e^k l_e^t}. \tag{19}$$

2) Bending:

The bending error $e_b(\mathcal{L}_k^t, \mathcal{T}_k)_n$ for node n connected with its neighbors $\boldsymbol{V}_l \in \mathcal{N}_j$ through the edge e_l is

$$e_b(\mathcal{L}_k^t, \mathcal{T}_k)_n = \frac{\delta_n^t - \delta_n^k}{l_{e_l}^k}. \tag{20}$$

where δ_n^t is the mean curvature of the surface at instant t. It is estimated through the neighbors of the node and itself.

$$\boldsymbol{\delta}_n^t = \boldsymbol{V}_n^t - \frac{1}{\sum_{l \in \mathcal{N}_j} \omega_l} \sum_{l \in \mathcal{N}_j} \omega_l \boldsymbol{V}_l^i \tag{21}$$

$$\delta_n^t = \|\boldsymbol{\delta}_n^t\|_2. \tag{22}$$

We assume fixed the values of the weights. Its derivative with respect to the node $\boldsymbol{V}_n^t$ is

$$\frac{\partial e_b(\mathcal{L}_k^t, \mathcal{T}_k)_n}{\partial \boldsymbol{V}_n^t} = \frac{\boldsymbol{\delta}_n^t}{l_{e_l}^k \delta_n^t} \tag{23}$$

with respect to its neighbors $\boldsymbol{V}_l^t$

$$\frac{\partial e_b(\mathcal{L}_k^t, \mathcal{T}_k)_n}{\partial \boldsymbol{V}_l^t} = \frac{\omega_l}{\sum_{l \in \mathcal{N}_j} \omega_l} \frac{\boldsymbol{\delta}_t^n}{l_{e_l}^k \delta_t^n}. \tag{24}$$

In case of being a plane, the mean curvature and its derivative tend to zero.

$$\frac{\partial e_b(\mathcal{L}_k^t, \mathcal{T}_k)_n}{\partial \boldsymbol{V}_n^t} = 0\,, \quad \boldsymbol{\delta}_n^t = 0. \tag{25}$$

3) Reference:

The reference error is

$$e_r(\mathcal{L}_k^t, \mathcal{L}_k^k) = \mathbf{V}_n^t - \mathbf{V}_n^k \tag{26}$$

and its derivative

$$\frac{\partial e_r(\mathcal{L}_k^t, \mathcal{T}_k)}{\partial \boldsymbol{V}_n^t} = 1. \tag{27}$$

REFERENCES

[1] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[2] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Int. Symp. Mixed Augmented Reality*, 2007, pp. 225–234.

[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[4] A. Chhatkuli, D. Pizarro, and A. Bartoli, "Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity," in *Proc. British Mach Vision Conf.*, 2014.

[5] A. Chhatkuli, D. Pizarro, T. Collins, and A. Bartoli, "Inextensible non-rigid shape-from-motion by second-order cone programming," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1719–1727.

[6] S. Parashar, D. Pizarro, and A. Bartoli, "Isometric non-rigid shape-from-motion with Riemannian geometry solved in linear time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2442–2454, Oct. 2018.

[7] J. Taylor, A. D. Jepson, and K. N. Kutulakos, "Non-rigid structure from locally-rigid motion," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 2761–2768.

[8] S. Vicente and L. Agapito, "Soft inextensibility constraints for template-free non-rigid reconstruction," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 426–440.

[9] A. Chhatkuli, D. Pizarro, A. Bartoli, and T. Collins, "A stable analytical framework for isometric shape-from-template by surface integration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 833–850, May 2017.

[10] J. Lamarca and J. M. M. Montiel, "Camera tracking for SLAM in deformable maps," in *Proc. 4th Inter. Workshop Recovering 6D Object Pose*, 2018.

[11] D. T. Ngo, J. Östlund, and P. Fua, "Template-based monocular 3D shape recovery using Laplacian meshes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 172–187, Jan. 2016.

[12] M. Salzmann and P. Fua, "Linear local models for monocular reconstruction of deformable surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 931–944, May 2011.

[13] P. Mountney, D. Stoyanov, and G.-Z. Yang, "Three-dimensional tissue deformation recovery and tracking," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 14–24, Jul. 2010.

[14] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake, "MIS-SLAM: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4068–4075, Oct. 2018.

[15] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 343–352.

[16] O. Sorkine and M. Alexa, "As-rigid-as-possible surface modeling," in *Eurographics*, vol. 4, 2007, pp. 109–116.

[17] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, "VolumeDeform: Real-time volumetric non-rigid reconstruction," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 362–379.

[18] W. Gao and R. Tedrake, "Surfelwarp: Efficient non-volumetric single view dynamic reconstruction," *arXiv preprint arXiv:1904.13073*, 2019.

[19] 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 Robot. Autom. Lett.*, vol. 3, no. 1, pp. 155–162, Jan. 2017.

[20] R. W. Sumner, J. Schmid, and M. Pauly, "Embedded deformation for shape manipulation," in *Proc. Assoc. Comput. Machinery's Special Interest Group Comput. Graphics Interactive Tech. Papers*, 2007, p. 80.

[21] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. Montiel, "Visual slam for handheld monocular endoscope," *IEEE Trans. Med. Imag.*, vol. 33, no. 1, pp. 135–146, Jan. 2014.

[22] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for hand-held monocular endoscopy," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 79–89, Jan. 2019.

[23] A. Marmol, A. Banach, and T. Peynot, "Dense-ArthroSLAM: Dense intra-articular 3d reconstruction with robust localization prior for arthroscopy," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 918–925, Apr. 2019.

[24] A. Bartoli, Y. Gérard, F. Chadebecq, T. Collins, and D. Pizarro, "Shape-from-template," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2099–2118, Oct. 2015.

[25] T. Collins and A. Bartoli, "Locally affine and planar deformable surface reconstruction from video," in *Proc. Int. Workshop Vision Model. Visualization*, 2010, pp. 339–346.

[26] A. Agudo, L. Agapito, B. Calvo, and J. M. M. Montiel, "Good vibrations: A modal analysis approach for sequential non-rigid structure from motion," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1558–1565.

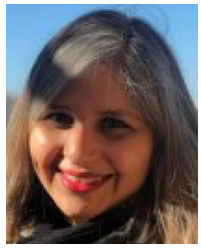
[27] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2000, pp. 690–696.

[28] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," in *Proc. Int. J. Comput. Vision*, vol. 107, no. 2, pp. 101–122, 2014.

[29] F. Moreno-Noguer and J. M. Porta, "Probabilistic simultaneous pose and non-rigid shape recovery," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 1289–1296.
[30] M. Paladini, A. Del Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito, "Factorization for non-rigid and articulated structure using metric projections," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 2898–2905.
[31] R. Garg, A. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 1272–1279.
[32] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Trajectory space: A dual representation for nonrigid structure from motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1442–1456, Jul. 2011.
[33] A. Agudo and F. Moreno-Noguer, "Simultaneous pose and non-rigid shape with particle dynamics," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2179–2187.
[34] P. F. Gotardo and A. M. Martinez, "Kernel non-rigid structure from motion," in *Proc. Int. Conf. Comput. Vision*, 2011, pp. 802–809.
[35] P. F. Gotardo and A. M. Martinez, "Non-rigid structure from motion with complementary rank-3 spaces," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 3065–3072.
[36] M. S. Floater, "Mean value coordinates," *Comput. Aided Geometric Des.*, vol. 20, no. 1, pp. 19–27, 2003.
[37] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proc. 9th IEEE Int. Conf. Comput. Vision*, 2003, p. 1403.
[38] D. Pizarro, R. Khan, and A. Bartoli, "Schwarps: Locally projective image warps based on 2d Schwarzian derivatives," in *Proc. Int. J. Comput. Vision*, vol. 119, no. 2, 2016, pp. 93–109.
[39] G. Bradski, "The OpenCV library," *Dr. Dobb's J. Softw. Tools*, 2000.
[40] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in *Proc. Int. Conf. Robot. Autom.*, 2011, pp. 3607–3613.
[41] S. Agarwal *et al.*, "Ceres solver," 2010. [Online]. Available: http://ceres-solver.org
[42] D. Stoyanov, G. P. Mylonas, F. Deligianni, A. Darzi, and G. Z. Yang, "Soft-tissue motion tracking and structure estimation for robotic assisted MIS procedures," in *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Springer, 2005, pp. 139–146.
[43] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G.-Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Springer, 2010, pp. 275–282.

Jose Lamarca received the B.S. and M.S. degrees in industrial engineering in 2014 and 2016, respectively, from Universidad de Zaragoza, Spain, where he is currently working toward the Ph.D. degree in the I3A robotics, perception, and real-time group.

His research interests include real-time visual SLAM for rigid and deformable environments.



Shaifali Parashar received the Ph.D. degree in computer vision from the Universite d'Auvergne, Clermont-Ferrand, France, in 2017.

She is currently a Postdoctoral Researcher with CVLab and EPFL in Lausanne, Switzerland. Her research interests include 3D computer vision including nonrigid 3D reconstruction and deformable SLAM.

Adrien Bartoli received the Ph.D. from the Perception Group, Inria Grenoble, Grenoble, France, in 2003, and the Habilitation degree from Université Blaise Pascal, Clermont-Ferrand, France, in 2008.

He has been a Professor of Computer Science with the Université Clermont Auvergne, Clermont-Ferrand, France, since fall 2009 and a member of Institut Universitaire de France from 2016 to 2021. He is currently leading the EnCoV (Endoscopy and Computer Vision) Research Group jointly with Michel Canis. He held an ERC Consolidator Grant (2013–2018) and an ERC Proof-of-Concept Grant (2018–2019). Previously, he was a CNRS Research Scientist with Institut Pascal where he led ComSee, the Computer Vision Research Group, jointly with Thierry Chateau. He was a Visiting Professor in DIKU at the University of Copenhagen between 2006 and 2009 and a Postdoctoral Researcher in the Visual Geometry Group with the University of Oxford under Andrew Zisserman in 2004. He has authored or coauthored approximately 100 scientific papers. His main research interests include image registration and Shape-from-X for rigid and nonrigid scenarios, and machine learning within the field of theoretical and medical computer vision.

Dr. Bartoli was the recipient of several awards including the 2004 Grenoble-INP Ph.D. thesis prize, the 2008 CNRS médaille de bronze, and the 2016 research prize from Université d'Auvergne. He has been on the program committees for top-ranking conferences in the field. He is on the editorial board of the *International Journal of Computer Vision* and *Journal of Artificial Intelligence Research* and was on the editorial board of IET CV and ELCVIA.

J. M. M. Montiel was born in Arnedo, Spain, in 1967. He received the M.S. and Ph.D. degrees in electrical engineering from Universidad de Zaragoza, Spain, in 1992 and 1996, respectively.

He is currently a Full Professor with the Departamento de Informtica e Ingeniera de Sistemas, Universidad de Zaragoza, where he is in charge of perception and computer vision research grants and courses. His interests include real-time visual SLAM for rigid and nonrigid environments, and the transference of this technology to robotic and nonrobotic application domains.

Prof. Montiel has received several awards including the IEEE Transactions on Robotics King-Sun Fu Memorial Best Paper Award 2016. Since 2020, he has been coordinating the EU FET EndoMapper grant aimed to bring visual SLAM to medical intracorporeal scenes. He has been awarded several Spanish MEC grants to fund research with the University of Oxford, U.K., and with Imperial College London, U.K.