Visual SLAM for Hand-Held Monocular Endoscope

Óscar G. Grasa, Ernesto Bernal, Santiago Casado, Ismael Gil and J. M. M. Montiel

***Abstract*—Simultaneous Localisation And Mapping (SLAM) methods provide real-time estimation of 3D models from the sole input of a hand-held camera, routinely in mobile robotics scenarios. Medical endoscopic sequences mimic a robotic scenario in which a hand-held camera (monocular endoscope) moves along an unknown trajectory while observing an unknown cavity. However, the feasibility and accuracy of SLAM methods have not been extensively validated with human in-vivo image sequences. In this work, we propose a monocular visual SLAM algorithm tailored to deal with medical image sequences in order to provide an up-to-scale 3D map of the observed cavity and the endoscope trajectory at frame rate. The algorithm is validated over synthetic data and human in-vivo sequences corresponding to fifteen laparoscopic hernioplasties where accurate ground-truth distances are available. It can be concluded that the proposed procedure is: 1) non invasive, because only a standard monocular endoscope and a surgical tool are used; 2) convenient, because only a hand-controlled exploratory motion is needed; 3) fast, because the algorithm provides the 3D map and the trajectory in real time; 4) accurate, because it has been validated with respect to ground-truth; and 5) robust to inter-patient variability, because it has performed successfully over the validation sequences.**

***Index Terms*—Endoscopy, Abdomen, Medical robotics, Virtual/augmented reality, Computer vision, SLAM**

I. Introduction

SLAM (Simultaneous Localisation And Mapping) is one of the most researched topics in mobile robotics: given a mobile sensor moving along an unknown trajectory in an unknown environment, the goal is to estimate, simultaneously, both the environment structure (a map of 3D points) and the sensor location with respect to that map. Only the information gathered by the sensor is taken as the input data to the algorithm; additionally, real-time performance at frame rate is a common requirement. The SLAM problem is particularly challenging in the case of monocular cameras because only a sequence of 2D projections of a 3D scene is available; in any case, 30 Hz real-time systems estimating up-to-scale 3D camera motions and maps of 3D points using commodity cameras and computers are widely available for mobile robotics environments nowadays. A seminal work by Davison [1] provided the first live working system. Being based on EKF (Extended Kalman Filter), its main weaknesses are the lack of robustness with respect to motion clutter and delayed feature initialisation. In [2], a robustified version of [1] is proposed by combining EKF with RANSAC. In [3], Klein and Murray propose a bundle-adjustment-based method that is currently one of the best performers in robotics. Previous works provide extensive experimental validation of the SLAM algorithms and its performance in real time for man-made, mainly rigid, scenes which are typical in mobile robotics. Recently, EKF based methods are being extended to provide estimates for non-rigidly deforming scenes in real time [4], [5].

Exploration of a body cavity with an endoscope can be posed as a monocular SLAM problem where an up-to-scale 3D map of the observed cavity is estimated from the sole input of an image sequence, gathered from a standard hand-held monocular endoscope, without resorting to any additional sensor such as optical or magnetic trackers, accelerometers, structured light, or artificial landmarks; the absolute scale of the map is recovered from the observation of a known size surgical tool. It is worth noting that the surgeon just needs to follow vague visibility-based directives about how to explore the cavity and it is not necessary to tightly follow a predefined endoscope motion. Furthermore, it is worth noting that SLAM not only recovers the 3D model, but also the actual trajectory followed by the endoscope. The map and trajectory estimates provide scene 3D measurements and support for augmented reality annotations.

This paper presents a monocular SLAM algorithm tailored to deal with medical endoscopic sequences and validated with in-vivo human medical sequences corresponding to ventral hernia repair surgeries. This work is the culmination of a series of previously published works. In [6], we proposed to use EKF + JCBB SLAM to process real hand-held monocular endoscopic sequences in order to measure distances or insert augmented reality annotations. Joint Compatibility Branch and Bound (JCBB) [7] is a state-of-the-art robust data association in EKF SLAM that exploits the correlation between pairs of matches to detect and reject spurious matches, however, its exponential computational complexity in the number of spurious hinders real-time performance even if there are more than one mismatch. We proposed a combination of EKF monocular SLAM + 1-point RANSAC ([2]) in [8] that results in an algorithm which is able to cope with high outlier rate in real time. Additionally, the Randomized List Relocalisation [9] is included to recover from endoscope tracking failure. In [10], we summarily proposed a hernia repair procedure based on visual SLAM. However, no experimental validation was provided.

In this work, an extensive validation of [8] over synthetic data and in-vivo sequences corresponding to fifteen real human ventral hernia repair surgeries is presented. We focus on

This work was supported by Spanish DPI2009-07130 and DPI2012-32168.

Ó. G. Grasa and J. M. M. Montiel are with Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain.{oscgg, josemari}@unizar.es.

E. Bernal, S. Casado and I. Gil are with Hospital Clínico Universitario "Lozano Blesa", Zaragoza, Spain.

An attached supplementary 3.8 MB video encoded with H.264/MPEG-4 AVC explaining the system behaviour is available.

geometrical accuracy providing a comparison with respect to ground-truth.

II. Related Work

A seminal work in providing 3D models from body monocular image sequences was proposed by Burschka *et al.* in [11]. Assuming scene rigidity, the system produces a map for registering preoperative CT scan with the endoscopic images. Its main limitations are map size and the lack of robustness with respect to outlier matches. Computer vision methods based on a discrete set of views have been applied to medical images, assuming scene rigidity, in order to just compute the 3D structure of the cavity. In [12], the classical two view RANSAC structure from motion is applied to mannequin images to determine the 3D structure; a constraint-based factorization 3D modelling method produces a dense 3D reconstruction in near real time. In [13], structure from motion is used to build a photorealistic 3D reconstruction of the colon; in a first stage, images are processed pairwise to produce an initial 3D map; in a second stage, all the maps are joined in a unique photorealistic 3D cavity model. In [14], these methods have been refined and extended to deal with multiple views with a significant boost in performance in rigid medical scenes. Thanks to careful feature selection and a quite robust spurious tracking ASKC [15] (Adaptive Scale Kernel Consensus), they are able to estimate both 3D models of the cavity and the location of the camera with respect to this cavity up to submilimeter accuracy in a cadaver for endonasal skull surgery. Hu *et al.* in [16], in a similar vein, propose a 3D structure estimation from multiple images. They deal with outliers by means of the trifocal tensor, then a bundle adjustment optimization is performed reporting accuracies slightly over a millimeter.

Cavity 3D reconstruction from medical sequences of non-moving stereo endoscope has been proposed in [17] and [18]. Visual SLAM methods have proven to be valid processing medical images coming from a moving stereo endoscope in [19], where an EKF stereo SLAM, assuming smooth camera motion and scene rigidity, is validated over synthetic sequences and qualitatively over in-vivo animal sequences; no usage of algorithms robust to spurious data is reported. In [20], the scene non-rigidity is considered: EKF visual SLAM is combined with a dynamic periodic model, learnt on-line, to estimate the respiration cycle from stereo images.

Intensive research is being done in designing medical miniaturised devices that can provide depth map as stereo endoscopes while avoiding the correspondence problem. A monoport structured light device based on a stereo scope is presented in [21], preliminary but promising results are reported. In [22], a catadioptric structured light prototype specifically designed to recover the lumen of a tubular cavity is described, reporting 0.1 mm accuracy tested on a phantom and ex-vivo animal. In [23], a monoport prototype combining time-of-flight (ToF) and RGB is proposed; despite the low resolution of the depth map, promising results are reported. All these previous devices are still under development, in any case the rich 3D information that they can provide suggests a promising venue of research for SLAM algorithms.

Our proposal is also based on EKF SLAM, however, we deal with monocular sequences, our method is robust to outliers, and we provide extensive validation over both synthetic data and real human in-vivo sequences.

Malti *et al.* in [24] propose a 2-phase 3D monocular reconstruction of the abdominal cavity using a template-based method. The first phase consists in exploring the abdominal cavity in order to obtain an initial 3D rigid reconstruction using 2 views and the essential matrix + camera resection + bundle adjustment combination. Afterwards, in the second phase, this reconstruction is exploited to infer 3D scene deformations during operation. The algorithm is one of the first to deal with the scene non-rigidity under general deformation. However, the correspondences are assumed known, computing time is not reported, and only a qualitative validation over one sequence is provided.

Recently, methods based on photometric properties are being used to endoscopic sequences. In [25], the results of [24] are taken as input to provide a dense 3D model based on shape from shading; only a quantitative validation for synthetic data, and qualitative validation for one in-vivo sequence of the uterus are provided. Collins *et al.* in [26] propose shape from shading in real time at 23Hz for medical rigid scenes thanks to a GPGPU implementation; in-vivo and ex-vivo experimental validation is provided but the authors acknowledge poor conditioning and the strong assumption of a constant albedo as prior data. [27] proposes a preliminary work based on photometric stereo with learnt reflectance models in order to estimate a 3D reconstruction of an organ from one image using 3 different color light sources; for this, the tip of the endoscope has to be modified to include 3 color filters. The method is able to compute the absolute depth without detecting image features, although it is sensitive to illumination changes. They provide preliminary experiments over one in-vivo pig liver sequence, including comparison with respect to ground-truth. The main advantage of photometric methods with respect to feature-based ones is their ability to deal with textureless images. However, they are still sensitive to illumination changes.

All the above mentioned methods are able to produce camera location with respect to the observed scene, a basic requirement for augmented reality insertions, navigation, or multimodal image fusion that have proven to be useful in medical applications for example [28], [29]. In [30], EKF stereo SLAM is also used to artificially expand the intraoperative field of view (FoV) of the laparoscope (dynamic view expansion).

Finally, it is worth mentioning the recent review about optical 3D reconstruction from medical image laparoscopic sequences provided by Maier-Hein *et al.* in [31].

III. Monocular Visual SLAM

This section is devoted to describing the EKF algorithm used in this work which is efficient as well as robust-to-spurious data association. We have selected the 1-point RANSAC (1PR) [2], in its exhaustive hypotheses testing version, as a robust-to-spurious data association method.

To compute the EKF estimation, it is mandatory to define the state, the state transition equations, and the measurement model. These definitions, for the case of the visual SLAM problem, are detailed in Sections III-A and III-B. The 1PR data association, like most of the EKF SLAM algorithms, computes firstly a putative individually compatible (IC) data association based on the EKF innovation computed at the prediction stage, it is detailed in Section III-C. Secondly, in a modified update stage, spurious matches are detected and removed, and only the inliers are eventually fused in the EKF (Section III-D). Next, the algorithm to add and remove features from the map (map management) is described in Section III-E. Finally, Section III-F describes the mandatory capability to relocalise the camera after tracking is lost.

A. State Vector Definition

The probabilistic representation of the world map and the camera location at step k is coded in a unique state vector modelled as a multivariate Gaussian, $\mathbf{x_k}$:

$$\mathbf{x}_k = \left(\mathbf{x}_v^\top, \mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_n^\top\right)^\top \quad . \tag{1}$$

It is composed of camera state, $\mathbf{x}_v$, and the map defined by location of every point, $\mathbf{y}_i$. See Section III-E for map point management details.

The camera state, $\mathbf{x}_v$, is formed from position, $\mathbf{r}^{WC}$, orientation encoded in a quaternion, $\mathbf{q}^{WC}$, and linear and angular velocities, $\mathbf{v}^W$ and ω^C.

The map is composed of n point features $(\mathbf{y}_1^\top \ldots \mathbf{y}_n^\top)^\top$ whose locations are encoded either in Euclidean coordinates, $\mathbf{y}_i = \begin{pmatrix} X_i & Y_i & Z_i \end{pmatrix}^\top$, or in inverse depth *(ID)*, $\mathbf{y}_i = \begin{pmatrix} x_i & y_i & z_i & \theta_i & \phi_i & \rho_i \end{pmatrix}^\top$. Details for the ID parametrization can be found in [32].

Regarding the state transition equation for the camera, we propose a dynamic constant velocity model to encode its smooth motion:

$$\mathbf{f}_v = \begin{pmatrix} \mathbf{r}_{k+1}^{WC} \\ \mathbf{q}_{k+1}^{WC} \\ \mathbf{v}_{k+1}^{W} \\ \omega_{k+1}^{C} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_{k}^{WC} + \left(\mathbf{v}_k^W + \mathbf{V}_k^W\right)\Delta t \\ \mathbf{q}_{k}^{WC} \times \mathbf{q}\left(\left(\omega_k^C + \Omega^C\right)\Delta t\right) \\ \mathbf{v}_k^W + \mathbf{V}^W \\ \omega_k^C + \Omega^C \end{pmatrix} \tag{2}$$

where $\mathbf{q}\left(\left(\omega_k^C + \Omega^C\right)\Delta t\right)$ is the quaternion defined by the rotation vector $\left(\omega_k^C + \Omega^C\right)\Delta t$.

We assume that the state noise vector is composed of linear, $\mathbf{a}^W$, and angular acceleration, α^C, acting as inputs producing, at each step, an impulse of linear velocity, $\mathbf{V}^W = \mathbf{a}^W \Delta t$, and angular velocity $\Omega^C = \alpha^C \Delta t$. Both of them are modelled as zero mean Gaussians with known covariance, $\mathrm{diag}\left(\mathbf{Q}_{\mathbf{a}^W}, \mathbf{Q}_{\alpha^C}\right)$, processes.

Regarding the state transition equation for the scene points, we propose a static model with zero state noise to encode the scene as perfectly rigid:

$$\mathbf{y}_{i_{k+1}} = \mathbf{y}_{i_k}. \tag{3}$$

B. Measurement Equation

The measurements, $\mathbf{z}_k = \mathbf{h}\left(\mathbf{x}_k\right)$, are provided by a pinhole camera:

$$\mathbf{h} = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u_0 - \frac{f}{d_x}\frac{h_x^C}{h_z^C} \\ v_0 - \frac{f}{d_y}\frac{h_y^C}{h_z^C} \end{pmatrix} \tag{4}$$

where u, v are the pixel coordinates of the observation in the image. $\mathbf{h^C} = \begin{pmatrix} h_x^C & h_y^C & h_z^C \end{pmatrix}^\top$, is the vector joining the current camera location with the observed map feature, expressed in the camera frame. u_0, v_0, f, d_x, d_y are the camera intrinsic parameters corresponding to the principal point, the focal length, and the pixel size. Finally, the two-parameter distortion model [33] is applied to compensate the lens radial distortion.

C. Classical EKF Estimation

Classical EKF estimation equations are:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{f}_k(\hat{\mathbf{x}}_{k-1|k-1}) \tag{5}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^\top + \mathbf{G}_k \mathbf{Q}_k \mathbf{G}_k^\top \tag{6}$$

$$\nu_k = \mathbf{z}_k - \mathbf{h}(\hat{\mathbf{x}}_{k|k-1}) \tag{7}$$

$$\mathbf{S}_k = \mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^\top + \mathbf{R}_k \tag{8}$$

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{H}_k^\top \mathbf{S}_k^{-1}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \nu_k \tag{9}$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{P}_{k|k-1} \tag{10}$$

where $\mathbf{F}_k = \frac{\partial \mathbf{f}_k}{\partial \mathbf{x}}$, $\mathbf{f}_k$ being the stacking of (2) and an instance of (3) for each map point. $\mathbf{G}_k = \frac{\partial \mathbf{f}_k}{\partial \mathbf{n}_k}$, where $\mathbf{n}_k$ is the state noise vector assumed to be a zero mean multivariate normal distribution with known covariance $\mathbf{Q}_k = \mathrm{diag}\left(\mathbf{Q}_{\mathbf{a}^W}, \mathbf{Q}_{\alpha^C}, 0^1, \ldots, 0^n\right)$, where each 0^i corresponds to a map point. $\mathbf{H}_k = \frac{\partial \mathbf{h}_k}{\partial \mathbf{x}}$. $\mathbf{R}_k$ is the image measurement error covariance, also assumed to be a diagonal matrix.

The first two equations (5, 6) encode the prediction step. The EKF prediction $\hat{\mathbf{x}}_{k|k-1}$ provides an estimate for the relative pose of every map point with respect to the camera. It is accurate enough to synthesize in a patch the point image appearance, compensating for rotation and scale variations along the sequence. Then, the synthesised patch is exhaustively searched inside the elliptical region defined by innovation and its covariance (7,8) by means of normalised image correlation (see Fig.1a). The pixel scoring highest, $\mathbf{z}_i$, if over a threshold, is selected as the match in the new image. This stage produces the set of putative IC matches:

$$\mathbf{z}_k^{IC} = \left(\mathbf{z}_1, \ldots, \mathbf{z}_{m_k}\right)^\top \tag{11}$$

corresponding to some of the visible map points.

Next stage is the update (9,10) where the information provided by the IC matches is fed in the estimation. This is the most expensive step in terms of computational cost. It has to be stressed that if one or more of the matches are spurious, the whole estimation process might become wrecked. Next section is devoted to detailing the robust-to-spurious update stage.

D. 1PR-EKF Update

The IC matching stage produces an initial set of putative matches, (11), where we can identify 3 subsets (see Fig.1c):

- *low-innovation inliers*.- Inlier matches whose innovation according to (7) is small when compared with the standard deviation of the image measurement error.
- *high-innovation inliers*.- Inlier matches whose innovation according to (7) is big when compared with the standard deviation of the image measurement error. They correspond to points whose location is quite uncertain, for example just initialised points. These observations are quite informative, hence valuable enough to be kept.
- *outliers*.- They are not jointly compatible with the rest of the IC matches. They have to be detected and excluded from the update.

The exhaustive hypotheses testing 1PR algorithm implements a modification of the EKF update stage which aims to detect and reject spurious matches. The algorithm has three main stages which are summarised in Algorithm 1. Fig. 1 illustrates the algorithm steps over a sample image.

The initial data are the IC matches (Fig. 1a). The first stage is *hypotheses generation and consensus*. It consists of a variation of the classical RANSAC algorithm [34] that exploits the EKF properties. For each match in $\mathbf{z}^{IC}$, a hypothesis is generated by integrating only the $\mathbf{z}_i$ measurement according to (9). It is worth noting that the expensive covariance update (10) is not applied. Assuming that all the correlated error is corrected by this integration, the innovation covariance (8) can be approximated as the measurement error $\mathbf{S}_k \approx \mathbf{R}_k$ covariance. Hence, a cheap χ^2 test can be applied to identify the support for the hypothesis. Due to this approximation, only low-innovation inliers are going to be included in the support.

The most supported hypothesis is considered as the consensus hypothesis and the corresponding supporting matches, $\mathbf{z}^{li\_inliers}$, are the definitive *low-innovation inliers* (Fig. 1b). Non-supporting matches, $\mathbf{z}^{nonsupport}$, are not definitively labelled as outliers because some of them can be *high-innovation inliers*.

The second stage, *high-innovation inliers rescue*, starts with a partial update using the low-innovation inliers, $\mathbf{z}^{li\_inliers}$, including both the estimation (9) and the covariance (10) updates. For each non-supporting match, a χ^2 test based on the new and more accurate innovation covariance (8) is applied, what allows to accept as high-innovation inliers, $\mathbf{z}^{hi\_inliers}$, some of the $\mathbf{z}^{nonsupport}$ (Fig. 1c).

Finally, in the *high-innovation inliers update* stage, the estimation is updated with the rescued high-innovation matches (Fig. 1d). Although the final number of detected spurious matches is rather low, their rejection is a must for performance.

Regarding the number of tested hypotheses, on the one hand, the cardinality of the IC set is low (in the order of tens). On the other hand, we are able to generate a hypothesis from just one measurement. As a result, the cardinality of the hypothesis set is the same as the IC set and we can afford to exhaustively test all the hypotheses. It is worth noting that we do not need to resort to cutting the complexity by random sampling. Although our hypotheses generation is not random, we still keep the RANSAC name because, in any case, our method is quite akin to the popular algorithm.

Algorithm 1 Exhaustive Hypotheses 1PR EKF-Update

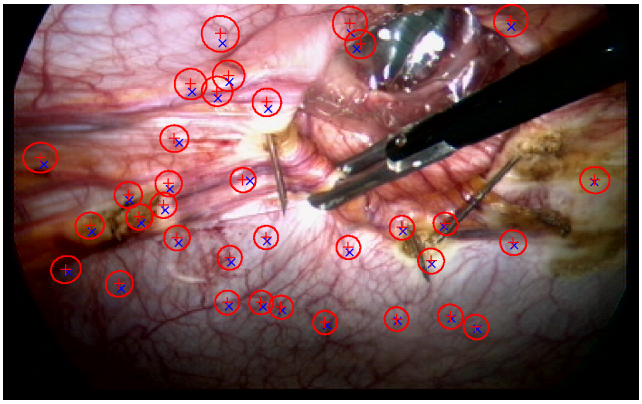
IN: $\hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1}$ {EKF prediction at step k}
  $\mathbf{z}^{IC}, \mathbf{R}_k$ {IC matches & Meas. Error Covariance}
OUT: $\hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}$ {EKF estimate at step k}
{*A. 1-Point hypotheses generation and consensus*}
for each $\mathbf{z}_i$ match in $\mathbf{z}^{IC}$ **do**
  $\hat{\mathbf{x}}_i = EKF\_state\_update(\mathbf{z}_i, \hat{\mathbf{x}}_{k|k-1})$(Eq. 9)
  $\hat{\mathbf{h}}_i = predict\_all\_measurements(\hat{\mathbf{x}}_i)$
  $[\mathbf{z}_i^{su}, \mathbf{z}_i^{ns}] = find\_supporters(\mathbf{z}^{IC}, \hat{\mathbf{h}}_i, \chi^2_{2,0.95}, \mathbf{R}_k)$
  if $size(\mathbf{z}_i^{su}) > size(\mathbf{z}^{li\_inliers})$ **then**
    $\mathbf{z}^{li\_inliers} = \mathbf{z}_i^{su}; \quad \mathbf{z}^{nonsupport} = \mathbf{z}_i^{ns}$
  end if
end for
{*B. Partial EKF update using low-innovation inliers & rescue high-innovation inliers*}
$[\hat{\mathbf{x}}_{k|k}^{li}, \mathbf{P}_{k|k}^{li}] = Update(\mathbf{z}^{li\_inliers}, \hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1})$(Eq. 9,10)
for each $\mathbf{z}^j$ match in $\mathbf{z}^{nonsupport}$ **do**
  $[\hat{\mathbf{h}}^j, \mathbf{S}^j] = point\_j\_pred\_and\_cov(\hat{\mathbf{x}}_{k|k}^{li}, \mathbf{P}_{k|k}^{li}, j)$
  $\nu^{\mathbf{j}} = \mathbf{z}^j - \hat{\mathbf{h}}^j$
  if $\nu^{\mathbf{j}\top}\mathbf{S}^{\mathbf{j}-\mathbf{1}}\nu^{\mathbf{j}} < \chi^{\mathbf{2}}_{\mathbf{2,0.95}}$ **then**
    $\mathbf{z}^{hi\_inliers} = add\_match\_to\_inliers(\mathbf{z}^{hi\_inliers}, \mathbf{z}^j)$
  end if
end for
{*C. Partial EKF update using high-innovation inliers*}
$[\hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}] = Update(\mathbf{z}^{hi\_inliers}, \hat{\mathbf{x}}_{k|k}^{li}, \mathbf{P}_{k|k}^{li})$(Eq. 9,10)

E. Map Management

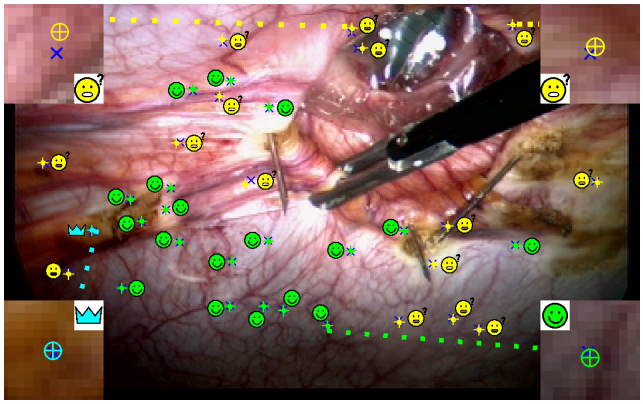
This section is devoted to describing how features are initialised and deleted from the map. As in any real-time visual SLAM method, an accurate prediction for the location of the map points with respect to the camera is available. This prediction allows to warp the patches defining the point appearance. Thanks to this warping, the perspective deformations are compensated. Then, resorting to expensive invariant descriptors and detectors would be overkilling in SLAM. Therefore, we propose to use FAST [35] and simple patch correlation to extract and recognise the map features (each map point is identified by a planar texture patch) because it is cheap and performs satisfactorily.

Additionally, in our current particular case, due to the small depth variation of the abdominal cavity and the limited laparoscope movements (it only pivots and slides over the fulcrum), features do not undergo severe perspective changes.

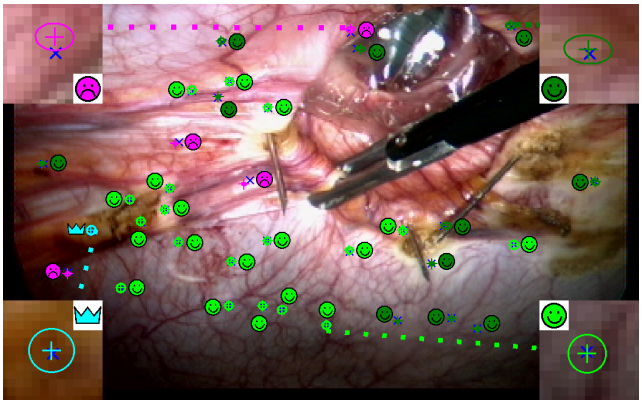
The feature initialisation criterion is targeted to keep in the FoV a predetermined number of visible features. When the number of visible features in the camera view is less than a threshold, features are initialised within a randomly located window favouring less populated areas (image regions with few or no map features). The strongest FAST corners are sought inside the window and, among them, the most distinctive one for relocalisation, as stated in Section III-F, is selected and initialised in the map. New features are encoded in ID and, as the estimation improves, converted to Euclidean.



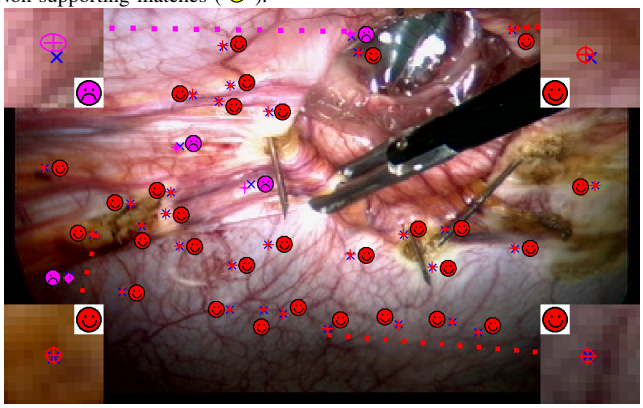
(a) Individually compatible –IC– matches. State prediction (+) with their corresponding elliptical search regions.



(b) Consensus hypothesis and low-innovation matches. The match generating the hypothesis ( ). Set of low-innovation supporting matches ( ). Non-supporting matches ( ).



(c) Low-innovation inliers update & high-innovation inliers rescue. Set of low-innovation inliers ( ). Rescued high-innovation inliers ( ) are now inside the search region and then accepted. Spurious matches ( ) remain out of the new search region.



(d) Final update. The updated state results from the integration of high and low-innovation inliers ( ). Outliers ( ) are not integrated.

Fig. 1. 1PR stages corresponding to one frame for the operation in Fig. 5a: (a) Individually compatible –IC– matches. (b) RANSAC winner hypothesis and consensus low-innovation matches. (c) Low-innovation partial update and the rescued high-innovation inliers. (d) Fully updated map. The estimated state is represented by its projection in the image, (+) stands for the estimation and the ellipse stands for the covariance. The measurements are displayed as (×). Different colors are used to code different matching categories. Zoom is made over 4 paradigmatic matches for each class matches.

A feature is removed from the map if it is repeatedly predicted to be in the image but it is not successfully matched. In our case, the reobservation rate should be higher than 40%. We can conclude that the surviving map features are trackable, locally salient, and distinctive for recognition at relocalisation.

F. Relocalisation

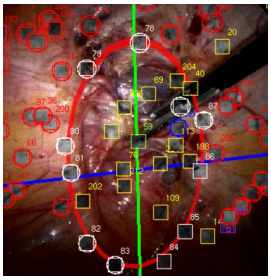
The previous active search strategy for matching is one of the system strengths (it enables the system to operate in real time) but it is also one of its weaknesses. The system works well provided that the mapped features are found inside the elliptical search regions. However, if the camera suffers sudden motions, the image is blurred, there are large occlusions, or the scene is deformed changing its appearance, the tracking will fail because no features are matched in several consecutive frames (Fig. 2).

The relocalisation algorithm must detect loss of tracking and stop EKF integration to avoid map corruption, due to incorrect data associations, and then enable a recovery procedure. If the tracking is lost, the relocalisation must find matches between the current image and the already estimated map in a data-driven manner without assuming priors about the camera location with respect to the map.

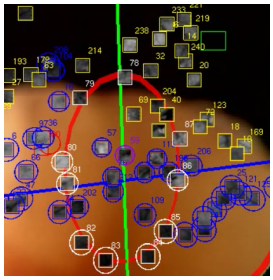
Our system uses Randomized List Relocalisation *(RLR)* proposed in [9]. It is summarised here with the aim of readability. RLR casts the image-to-map matching as a classification problem. A two stage online training is applied for every map feature. First, at feature initialisation, 400 warped versions of the texture patch around the feature are GPU-synthesised from the image where the feature is first observed. The warped patches are used to train the classifier. The second stage harvests texture patches during EKF operation that are used for online training.

The classifier is also exploited for selecting the most distinctive features at initialisation: only features scoring low in the classifier with respect to other features already in the map are eventually initialised.

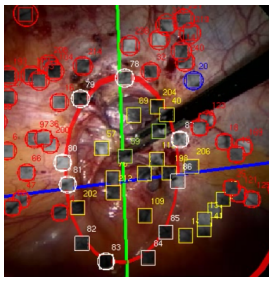
When the system detects a tracking failure (i.e. there are no observations in a frame, the camera pose uncertainty has grown too large, or the predicted camera view does not contain any mapped features) a few thousand of the strongest FAST features [35] detected in the current image are fed to the classifier to find putative image-to-map matches. As



(a) The system is on track.



(b) A large occlusion causes the tracking loss.



(c) Finally, the system relocalises and tracking is recovered.

Fig. 2. Relocalisation example for the operation in Fig. 5c.

map features can be similar to each other, multiple feature correspondence hypotheses are considered. Then RANSAC is applied to relocalise the camera with respect to the map. Camera location is hypothesized from three feature correspondences using three-point-pose PnP algorithm proposed in [34]. Each camera location hypothesis is rated according to how many other map features can be matched in the image. The consensus hypothesis is optimised in a "moving camera observing a fixed map" manner.

IV. Experimental Validation Description

The goal of the experimental validation[1] is to prove the feasibility of using monocular visual SLAM in real surgical procedures. We have selected ventral hernia repair as a paradigmatic example because:

1) The scene is almost rigid and textured.
2) The standard procedure already includes accurate distance measurements that can be used as ground-truth to assess the visual SLAM geometrical accuracy.
3) The flexibility and robustness of visual SLAM methods are clearly tested because the surgical procedure has not been modified at all, except for the addition of an exploratory endoscope manoeuvre with a trajectory similar to other endoscope routine motions.
4) The SLAM version, just by making better use of the images, would simplify the surgical procedure without a disruptive modification of the workflow.
5) The image sequences exhibit significant inter-patient variability in texture, illumination, input port placement, and exploratory trajectory.

[1]The experiments developed in this work were approved by Comité Ético de Investigación Clínica de Aragón (CEICA) and governed according to the provisions of the Spanish Law *14/2007* regarding biomedical research.

Fifteen in-vivo human laparoscopic ventral hernia repair interventions were captured at 384x288@25 fps with an optics with 30º direction of view (DoV) and 60º FoV angles. The standard procedure has been extended with the additional exploratory endoscope manoeuvre. The goal of the exploration is to gather a sequence detecting parallax for an accurate visual SLAM. At the end of each operation, a calibration planar pattern was imaged for camera calibration according to Zhang's method [36]. For twelve of the operations, it was possible to take tape measurements for, at least, one main axis of the hernia (ground-truth) (Fig. 3b). The reasons for not taking some of the measurements were the difficulty in manoeuvrability or surgical time saving due to some patients' medical conditions.

Additionally, a representative simulation is designed in order to quantitatively evaluate the accuracy and robustness of the method. The simulation mocks up the 3D geometry of the ventral hernia repair procedure where the human torso is modelled by means of an array of points on an ellipsoidal cap (Fig. 4f). Typical local non-rigid deformations of hernia repair emulating external forces, respiration or heartbeats have been applied over the cap. In the left flank of the cap, a virtual 30º DoV and 60º FoV endoscope and a virtual tool tip have been inserted. From this setup, a synthetic image sequence is generated by moving the virtual endoscope around the fulcrum mimicking the real laparoscope movements. The 3D model points are projected according to the pinhole + two-radial-distortion parameter model and adding zero mean Gaussian noise with 0.5 pixels standard deviation. It has been simulated not only at the actual endoscope resolution 384×288 pixels but also double 768×576 and half 192×144.

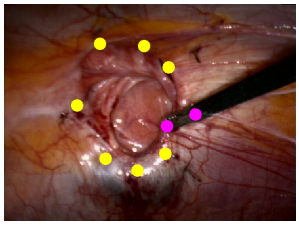
A. Classical Hernia Repair Procedure

The hernia defect is measured in-vivo to cover the defect with a customised-in-size patch. The elliptical patch axes are those of the defect plus a predefined safety margin. If possible, a piece of a sterilised tape measure is introduced inside the abdominal cavity and at least one of the two main hernia axes is measured (Fig. 3b). When available, we will use this measurement as ground-truth to validate the SLAM geometrical accuracy. The 0.5 cm tape measurement resolution determines the ground-truth accuracy. If the tape measurement cannot be taken other less accurate indirect methods are used. We opt not to consider them as ground-truth.

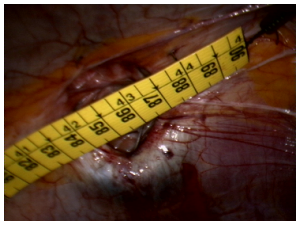
B. Hernia Repair SLAM Assisted Procedure

In addition to the standard procedure, at the measurement stage, an exploratory laparoscope manoeuvre is performed aimed at translating the endoscope tip while the region of interest is kept in the Fov (Fig. 3c). The sequence is processed to estimate a cavity map (cavity 3D model up to a scale factor) and the endoscope trajectory.

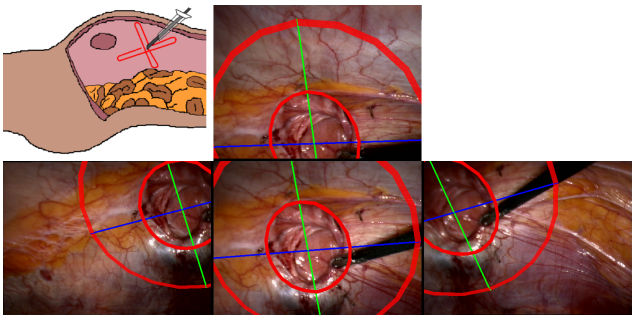
Before the exploratory manoeuvre, additional key points are manually enforced to be in the map: two predefined points over a clinch to define the scale, s, and several points (five or more) scattered over the defect boundary to estimate the hernia contour and size (Fig. 3a).



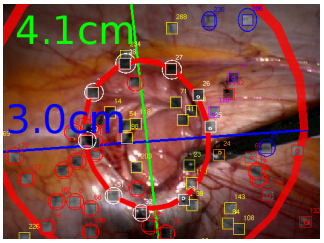
(a) Two points over a clinch define the scale (magenta). Five or more points over the hernia defect boundary (yellow).



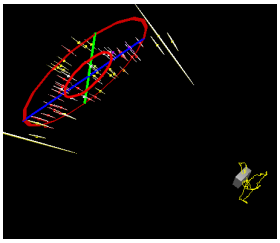
(b) Tape measurement considered as ground-truth.



(c) Hand-held exploratory laparoscope motion.



(d) SLAM measurement, Map and ellipses projected as augmented reality over a sequence frame.



(e) Camera trajectory, 3D map and ellipses in 3D. Top view.

Fig. 3. Measurement processes, 3D map, camera trajectory and the estimated ellipses for the operation in Fig. 5b.

The hernia defect is modelled as a virtual 3D ellipse in a three-stage way. In the first stage, an initial guess of the dominant plane defined by the five or more defect boundary points is computed by least squares. This guess is covariance-weighted in the second stage by an information filter extracting the needed covariances from the probabilistic map of the EKF monocular SLAM. After that, the points are projected on the weighted plane where the planar ellipse is fitted. Finally, the ellipse dimensions are computed from the scale factor s.

Resulting from the exploration, the SLAM algorithm estimates the scene 3D map and the endoscope trajectory (Fig. 3e). The defect major and minor axes sizes are estimated from the ellipse (Fig. 3d). A second concentric ellipse defining the virtual border of the patch is visualised as an augmented reality annotation. Live augmented reality is possible due to the real-time 3D estimation of the endoscope position with respect to the 3D cavity (Fig. 3d).

V. Experimental Results

This section details both the simulation and the real operation results. The same parameters, experimentally tuned, have been applied for all of the experiments: image measurement error of 0.5 pixels standard deviation; 40% is the acceptance threshold for normalised correlation score to eventually accept a map point match in the new image; new features are assigned an initial 1 inverse depth, with an initial $\sigma_\rho = 1$, in order to have an initial direct depth acceptance region starting in 0.3 and extending to include infinite; regarding linear and angular accelerations, standard deviations are $2.5\frac{1}{s^2}$ and $3\frac{rad}{s^2}$ respectively, as monocular cannot observe the scale, both depth and linear acceleration have no length units; finally, map management initialises features in order to have 40 map points observable in the image.

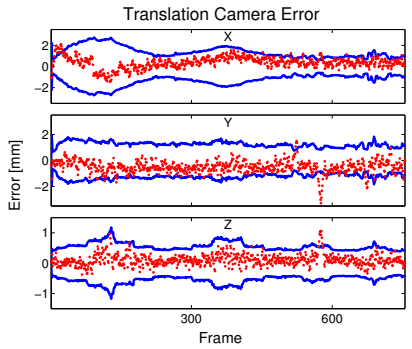
A. Simulation

We focus on the 384×288 resolution because it corresponds to the endoscope used in our surgeries. Figs. 4a and 4b display the estimation error history for the camera translation and rotation respectively. Both the error and the $\pm3\sigma$ acceptance region are represented. We can conclude that the EKF provides a consistent estimation because the estimated value is mostly within the 3σ interval. Additionally, thanks to the covariance estimation, we can evaluate how accurate the available estimation is at a given time step. The time evolution shows how initially the covariance grows due to the exploratory motion that departs from the initial camera location. As the estimation evolves, some features are reobserved and then the estimation error decreases.

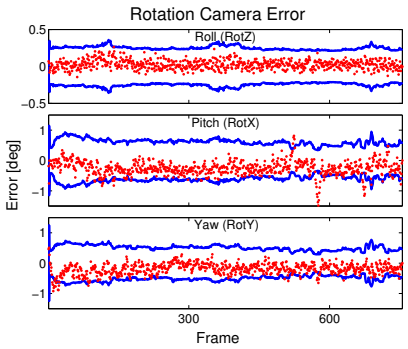
Fig. 4c displays the estimation error distributions for the camera estimation history by means of box-and-whisker diagrams. The left and right of the box represent the first and third quartiles, the line inside the box is the median; the ends of the whiskers represent the minimum and maximum of all of the data. The errors are in the interval [0.6, 1.1] mm with 0.82 mm as the median for translation, and [0.27, 0.49] deg with 0.38 deg as the median for rotation.

The estimated map corresponds to the "rigid envelope" where none of the points in the cap are deformed. During the simulation, observations corresponding to non-rigid deformations are successfully marked as spurious by 1PR and are not considered in the estimation. It is worth noting that if 1PR is disabled, some spurious matches are marked as inliers and the estimation fails. For each time step and for each map point, the EKF provides both an estimate for the location and its covariance. As more images are processed, the covariance for a given point is reduced if the point is reobserved. Fig. 4f displays the estimated map with the corresponding ellipsoidal 3σ acceptance regions after processing the whole sequence. It can be seen that most of the points have a small error except those at the map boundaries. Points on the boundary are only detected in a few images providing little parallax, hence their location error is great. In any case, we have verified that the estimation error normalised with the estimated covariance approximately distributes as a χ^2 with 3 d.o.f.(see Fig. 4e). We can conclude that the map estimation is consistent, hence estimated covariances provide a per point accuracy measurement.

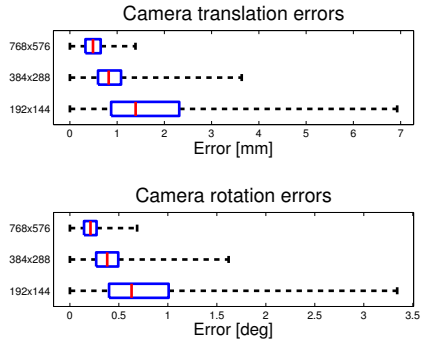
Fig. 4d displays box-and-whisker diagrams for the estimation error for all the map points after processing the whole



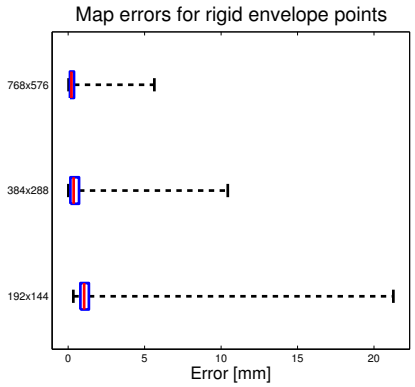
(a) 384x288 Camera translation error (red) and the corresponding $\pm 3\sigma$ acceptance region (blue).



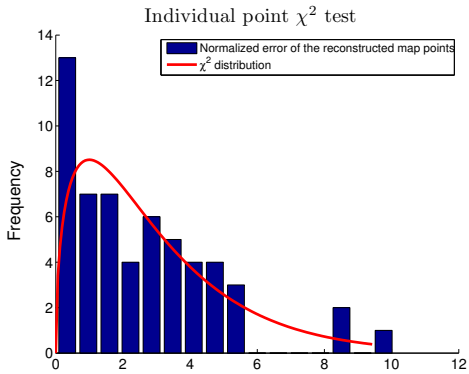
(b) 384x288 Camera rotation error (red) and the corresponding $\pm 3\sigma$ acceptance region (blue).



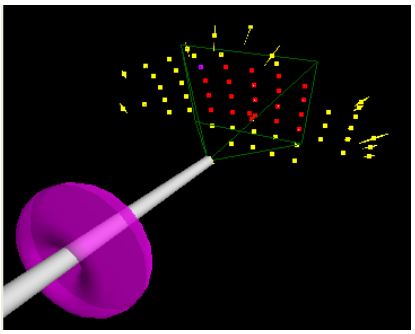
(c) Box-and-whisker diagrams for the camera location error in the three analysed resolutions.



(d) Box-and-whisker diagrams for the error of the map points in the three analysed resolutions.



(e) Map point error normalised with respect to the estimated covariance. It approximately distributes as the theoretical χ^2 with 3 d.o.f.



(f) Estimated map points and their covariances after processing the whole 384×288 sequence. The laparoscope (grey) is passing though the ground-truth fulcrum (magenta).

Fig. 4. Estimation error for the simulation results.

sequence. The errors are in the interval [0.15, 0.71] mm with 0.36 mm as the median, the maximum error is 10.44 mm corresponding to a point on the boundary.

From the 384×288 simulation, we can conclude that the the map estimation is accurate up to 1 mm for most of the points, in any case, the estimated covariance provides an assessment for each point accuracy. Regarding the effect of the camera resolution, the half and double resolution simulations show that EKF can make the most of the available resolution because error increases inversely with respect to the resolution (Figs. 4c and 4d).

B. Laparoscopic Sequences

Our EKF SLAM has been able to successfully compute the map and the camera trajectory for the fifteen sequences, see Fig. 5 for a thumbnail of all the surgeries. It has been able to cope with a variety of illuminations, textures and input port geometries. If we have to mention a weakness, it is the inability to perform the measurement in one of the sequences (Fig. 5o) because of the lack of texture around the defect.

To analyse the cycle time budget, we have selected the sequence corresponding to Fig. 5c because it is archetypical. It includes EKF routine operation and relocalisation after track loss due to occlusion (Fig. 2). Fig. 6a displays the cycle time budget split in: EKF prediction, putative IC matching, 1PR hypotheses generation and consensus, low innovation inliers update, high-innovation rescue and update, and map management (feature creation and removal). IC matching is time consuming due to image correlation and patch warping.

Fig. 6b displays the cost per frame histogram for all frames in all sequences (6473 frames). The cycle time mode is around 13 ms, the mean and the median being around 18 ms. Faster frames ($<$10 ms) correspond to relocalisation when no features are detected (since in that case no relocalisation hypotheses are generated), and with first sequence frames when the map is small. Times around 38 ms correspond to frames when the system has just relocalised and the camera location is still not refined. Thus, we can conclude that robust real time performance can be achieved.

Typical map sizes are between 50 and 100 points. Up to 40 map features are measured per frame. Fig. 6c shows a histogram of the outlier count for all frames in all the sequences. Although nearly 30% of frames do not contain any spurious match, only one of the sequences can be successfully processed if 1PR is disabled. We can thus conclude that algorithms robust to spurious data are a must for EKF SLAM even in the case of a low spurious-matches rate. 1PR cost is linear in number of measurements and state size while the outliers have a low influence on the computational cost ($<$20% of the total budget corresponding to 1PR hypotheses generation and consensus). Hence, our system can achieve real time even when $\sim$25% of frames contain more than 3 outliers.

In contrast, methods like exhaustive JCBB, with exponential complexity in the number of outliers, would not perform in real time.

Regarding the clinical point of view, we have to stress that, except for one case (Fig. 5o), we were always able to measure both ellipse axes because the defect visibility is required during the surgery and we profit from that. In the failing case, we were able to build the map, however, the clicked points signaling the defect were not trackable due to the lack of stable texture in the defect boundary area and the particular point detection method. A more dedicated work in image processing (e.g. using contours) is quite likely to overcome this limitation. In contrast, classical tape measuring procedure sometimes fails to produce the measurement because of the limited manoeuvrability resulting from the port placement.

The surgical time consumed by SLAM is mainly due to the exploration, which takes less than 1 minute irrespective of the sequence. Since the algorithm runs live (Fig. 6f), no additional time is needed for the processing, except for selecting the points over defect boundary and over the clinch to define the scale. Both are easy to automate with the corresponding surgical time saving. In contrast, the classical measurement procedure is rather uncertain (the time length ranges from 2 to 5 minutes). It has to be noted that in three cases where longer times were anticipated, the surgeons did not even try to measure. In these cases, other less accurate methods were used. In any case, SLAM recovers not only two measurements but a full 3D model and the support for augmented reality.

To validate the SLAM geometrical accuracy, the dimensions of the hernia defect's main axes have been estimated from the 3D recovered model and compared with those of tape measurement (our ground-truth), accurate up to 0.5cm. No significant differences can be observed so we can conclude that SLAM is as accurate as the tape measurement. Figs. 6d-6e depict measurements in the two axes.

VI. Clinical Assessment of the Method

The method can be easily integrated in the surgical workflow because it only needs standard elements operated in a standard manner. It does not need the insertion of any additional element but just a standard tool that has to be inserted in any case. It is able to produce accurate measurement at shorter surgical time than the other competing methods.

The tape measurement sometimes cannot be used because of the difficult manoeuvrability. Several needles insertion defining the defect mayor and minor axes is a common alternative, it is less accurate and has several additional shortcomings. It is invasive, there is risk of haemorrhage if a blood vessel is reached, if there is a previous surgery prior scar tissue can come in contact with the needle increasing the risk of infection and inflammation.

It can be easily extended to other surgical procedures such as thoracoscopy or flexible endoscopy. It can also be relevant in excisions for intraabdominal measurements of organs such as liver, adrenal glands, or spleen aiming to determine the size of the extraction port.

It is worth noting that in addition to the measurements, visual SLAM can provide support for augmented reality annotations of intraoperative information during the actual surgical procedure in real time, for example virtual landmarks for aligning the prosthetic patch.

VII. Conclusions

Traditional endoscope surgery displays and disposes of the image sequence. However, monocular SLAM is able to exploit the sequence.

We have provided the first human in-vivo experimental validation for the feasibility of using EKF monocular SLAM as a proper method to deal with medical endoscope sequences. A scene rigid model is assumed, however, the method has proven robust with respect to scene local non-rigid deformations such as respiration or external forces. The validation is based on synthetic data and on sequences coming from a real surgical environment over fifteen human in-vivo laparoscopic ventral hernia repair surgeries.

The method has proved to be fast, non-invasive, and easily incorporated to the existing surgical workflow by using solely images gathered from a hand-held standard monocular endoscope and standard laparoscopic tools.

Unlike other experimental validations based on phantoms or animal imagery, we have tested the fifteen human surgeries that displayed the typical inter-patient variability (different textures, illumination, input port placement and exploratory trajectories) (see Fig. 5 and the accompanying video). Despite the variability, all the sequences have been processed with the same tuning, therefore we believe that they provide experimental evidence of the method's usability.

We have tested the performance of the EKF monocular SLAM + 1PR, in any case, any real-time visual SLAM method, either monocular or stereo, would perform equally well on condition that it has a robust-to-spurious policy.

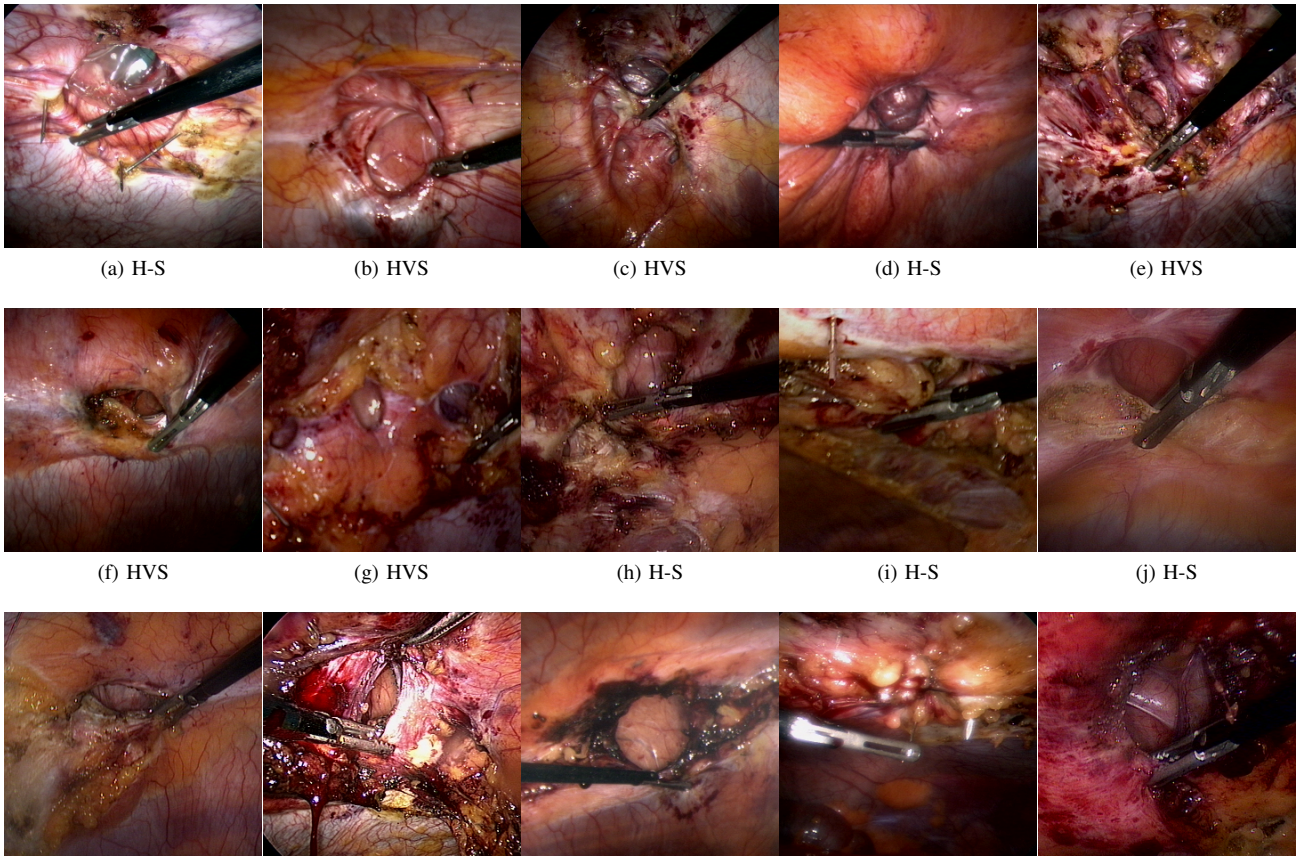
VIII. Future Work

This method cannot deal with non-rigid nor with textureless scenes. Besides, offline camera calibration is required.

Regarding the non-rigidity, Agudo *et al.* [4], [5] have proved that the combination EKF-FEM (Finite Elements Method) can deal with deformations in real time. This approach is relevant for medical images because it can exploit the biomechanical availability. One of our immediate goals is to adapt these methods to our system. Concerning calibration, it would be desirable to solve the complete problem (3D structure recovery, camera location and camera calibration) during the exploratory movement. Finally, the lack of texture could be tackled using a monocular SLAM based on points and edges and researching the combination with photometric methods.

In the particular case of the hernia measurement, another minor limitation is that currently the scale and the hernia defect have to be selected by clicking on the images; an automatic detection of both would ease the use of the system.

Since our system is based on an EKF implementation, we can only handle a few hundred points. However, methods based on keyframe + bundle adjustment such as [3] can render a map composed of a few thousand points. This signals a clear way for increasing the map density.



(a) H-S (b) HVS (c) HVS (d) H-S (e) HVS

(f) HVS (g) HVS (h) H-S (i) H-S (j) H-S

(k) HVS (l) –S (m) –S (n) –S (o) HV-

Fig. 5. The thumbnails –labelled from (a) to (o)– corresponding to the 15 ventral hernia repair surgeries used to validate the system. The "HVS" code in the captions stands for the availability of (H) Horizontal tape measurement, (V) Vertical tape measurement and (S) SLAM measurement. The SLAM map was successfully computed for all of them, while ellipse measurement was not possible in (o) due to the lack of texture around the defect.

Finally, a more ambitious goal is to exploit the camera location as a backbone for augmented reality providing additional visual information (multimodal registration images –CT or MRI– or another kind of annotations) in real time.

ACKNOWLEDGEMENT

The authors would like to thank Imperial College London (Dr. A.J. Davison) for collaboration in visual SLAM software; and University of Oxford (Dr. I. Reid and Dr. B. Williams) for visual SLAM and relocalisation software.

REFERENCES

[1] A. Davison, "Real-Time Simultaneous Localisation and Mapping with a Single Camera," in *ICCV*, 2003, pp. 1403–1410 vol.2.

[2] J. Civera, O. G. Grasa, A. J. Davison, and J. M. M. Montiel, "1-Point RANSAC for Extended Kalman Filtering: Application to Real-Time Structure from Motion and Visual Odometry," *Journal of Field Robotics*, vol. 27, no. 5, pp. 609–631, Sep. 2010.

[3] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," in *ISMAR*, 2007, pp. 225–234.

[4] A. Agudo, B. Calvo, and J. Montiel, "Finite Element based Sequential Bayesian Non-Rigid Structure from Motion," in *CVPR*, 2012, pp. 1418–1425.

[5] ——, "3D Reconstruction of Non-Rigid Surfaces in Real-Time Using Wedge Elements," in *5th Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment (ECCV)*, vol. 7583, 2012, pp. 113–122.

[6] O. G. Grasa, J. Civera, A. Güemes, V. Muñoz, and J. M. M. Montiel, "EKF Monocular SLAM 3D Modeling, Measuring and Augmented Reality from Endoscope Image Sequences," in *5th Workshop on Augmented Environments for Medical Imaging including Augmented Reality in Computer-Aided Surgery. (MICCAI)*, 2009.

[7] J. Neira and J. Tardós, "Data Association in Stochastic Mapping Using the Joint Compatibility Test," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 6, pp. 890–897, 2001.

[8] O. Grasa, J. Civera, and J. M. M. Montiel, "EKF Monocular SLAM with Relocalization for Laparoscopic Sequences," in *ICRA*, 2011, pp. 4816–4821.

[9] B. Williams, G. Klein, and I. Reid, "Real-Time SLAM Relocalisation," in *ICCV*, 2007, pp. 1–8.

[10] I. Gil, J. Marín, J. M. Martínez, E. Bernal, S. Casado, O. García, and J. Quintana, "Augmented Reality and 3D Measurement for Monocular Laparoscopic Abdominal Wall Hernia Repair," in *46th Congress of the European Society for Surgical Research (ESSR11)*, 2011.

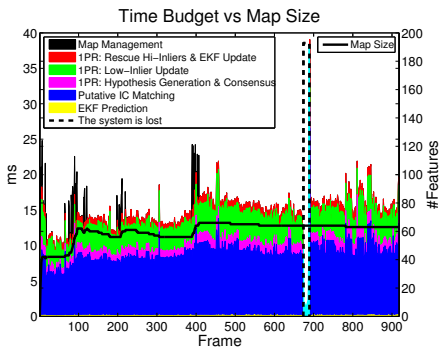
[11] D. Burschka, M. Li, M. Ishii, R. H. Taylor, and G. D. Hager, "Scale-Invariant Registration of Monocular Endoscopic Images to CT-Scans for Sinus Surgery," *Medical Image Analysis*, vol. 9, no. 5, pp. 413–426, 2005.

[12] C.-H. Wu, Y.-N. Sun, and C.-C. Chang, "Three-Dimensional Modeling From Endoscopic Video Using Geometric Constraints Via Feature Positioning," *IEEE Trans. on Biomedical Engineering*, vol. 54, no. 7, pp. 1199–1211, 2007.

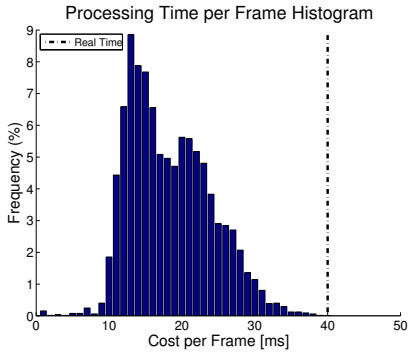
[13] D. Koppel, C.-I. Chen, Y.-F. Wang, H. Lee, J. Gu, A. Poirson, and R. Wolters, "Toward automated model building from video in computer-assisted diagnoses in colonoscopy," in *Proc. of the SPIE Medical Imaging Conf.*, 2007.

[14] D. Mirota, H. Wang, R. H. Taylor, M. Ishii, G. L. Gallia, and G. D. Hager, "A System for Video-Based Navigation for Endoscopic Endonasal Skull Base Surgery," *TMI*, vol. 31, no. 4, pp. 963–976, 2012.

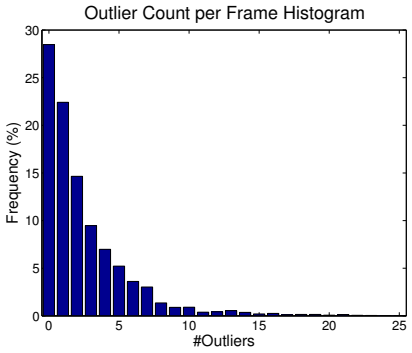
[15] H. Wang, D. Mirota, M. Ishii, and G. D. Hager, "Robust Motion Estimation and Structure Recovery from Endoscopic Image Sequences



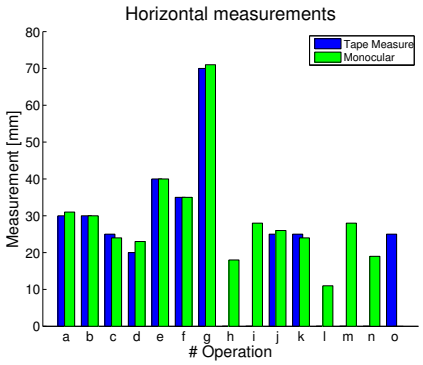
(a) Cycle time split in the six main parts of the EKF cycle and map size for an archetypal execution corresponding to operation Fig. 5c.



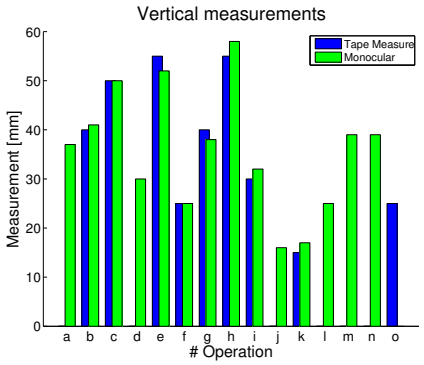
(b) Cycle time histogram for all processed frames in the 15 sequences.



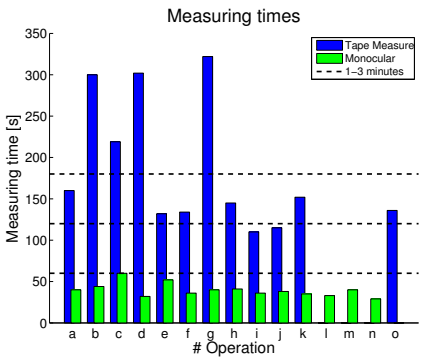
(c) Outlier histogram for all processed frames in the 15 sequences.



(d) Horizontal axis measurement comparison.



(e) Vertical axis measurement comparison.



(f) Measuring time comparison.

Fig. 6. (a) Cycle time and map size corresponding to operation Fig. 5c. (b) Cycle time and (c) outlier histograms for all frames in all sequences. (d), (e), (f) Measurement procedure comparison. Both accuracy (d), (e) and surgical time (f) are exhaustively plotted, one bar per operation per method. Missing data are represented as a missing bar. The labels correspond with those on Fig. 5.

with an Adaptive Scale Kernel Consensus Estimator," in *CVPR*, 2008, pp. 1–7.

[16] M. Hu, G. Penney, M. Figl, P. Edwards, F. Bello, R. Casula, D. Rueckert, and D. Hawkes, "Reconstruction of a 3D surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes," *Medical Image Analysis*, vol. 16, no. 3, pp. 597–611, 2012.

[17] F. Mourgues, F. Devernay, and É. Coste-Manière, "3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery," in *IEEE/ACM Symp. on Augmented Reality*, 2001, pp. 191–192.

[18] D. Stoyanov, A. Darzi, and G.-Z. Yang, "A Practical Approach Towards Accurate Dense 3D Depth Recovery for Robotic Laparoscopic Surgery," *Computer Aided Surgery*, vol. 10, no. 4, pp. 199–208, 2005.

[19] P. Mountney, D. Stoyanov, A. Davison, and G.-Z. Yang, "Simultaneous Stereoscope Localization and Soft-Tissue Mapping for Minimal Invasive Surgery," in *MICCAI*, vol. 4190, 2006, pp. 347–354.

[20] P. Mountney and G.-Z. Yang, "Motion Compensated SLAM for Image Guided Surgery," in *MICCAI*, vol. 6362, 2010, pp. 496–504.

[21] X. Maurice, C. Albitar, C. Doignon, and M. de Mathelin, "A structured light-based laparoscope with real-time organs' surface reconstruction for minimally invasive surgery," in *EMBC*, 2012, pp. 5769–5772.

[22] C. Schmalz, F. Forster, A. Schick, and E. Angelopoulou, "An endoscopic 3D scanner based on structured light," *Medical Image Analysis*, vol. 16, no. 5, pp. 1063–1072, 2012.

[23] S. Haase, C. Forman, T. Kilgus, R. Bammer, L. Maier-Hein, and J. Hornegger, "ToF/RGB Sensor Fusion for 3-D Endoscopy," *Current Medical Imaging Reviews*, vol. 9, no. 2, pp. 113–119, 2013.

[24] A. Malti, A. Bartoli, and T. Collins, "Template-Based Conformal Shape-from-Motion from Registered Laparoscopic Images," in *MIUA*, 2011.

[25] ——, "Template-Based Conformal Shape-from-Motion-and-Shading for Laparoscopy," in *IPCAI*, vol. 7330, 2012, pp. 1–10.

[26] T. Collins and A. Bartoli, "Towards Live Monocular 3D Laparoscopy Using Shading and Specularity Information," in *IPCAI*, vol. 7330, 2012, pp. 11–21.

[27] ——, "3D Reconstruction in Laparoscopy with Close-Range Photometric Stereo," in *MICCAI*, vol. 7511, 2012, pp. 634–642.

[28] A. Okur, S.-A. Ahmadi, A. Bigdelou, T. Wendler, and N. Navab, "MR in OR: First analysis of AR/VR visualization in 100 intra-operative Freehand SPECT acquisitions," in *ISMAR*, Oct. 2011, pp. 211–218.

[29] S. Nicolau, L. Soler, D. Mutter, and J. Marescaux, "Augmented reality in laparoscopic surgical oncology," *Surgical Oncology*, vol. 20, no. 3, pp. 189–201, 2011.

[30] J. Totz, P. Mountney, D. Stoyanov, and G.-Z. Yang, "Dense Surface Reconstruction for Enhanced Navigation in MIS," in *MICCAI*, vol. 6891, 2011, pp. 89–96.

[31] L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov, "Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery," *Medical Image Analysis*, vol. 17, no. 8, pp. 974–996, 2013.

[32] J. Civera, A. J. Davison, and J. M. M. Montiel, "Inverse Depth Parametrization for Monocular SLAM," *IEEE Transactions on Robotics (T-RO)*, vol. 24, no. 5, pp. 932–945, 2008.

[33] E. M. Mikhail, J. S. Bethel, and J. C. McGlone, *Introduction to Modern Photogrammetry*. John Wiley & Sons, 2001.

[34] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applicatlons to Image Analysis and Automated Cartography," *Com. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[35] E. Rosten and T. Drummond, "Fusing Points and Lines for High Performance Tracking," in *ICCV*, vol. 2, Oct. 2005, pp. 1508–1515.

[36] Z. Zhang, "Flexible Camera Calibration by Viewing a Plane From Unknown Orientations," in *ICCV*, vol. 1, 1999, pp. 666–673.