2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

Deformable Structure From Motion by Fusing Visual and Inertial Measurement Data

Stamatia Giannarou, Zhiqiang Zhang and Guang-Zhong Yang, *Fellow,IEEE*

Abstract— **Accurate recovery of the 3D structure of a deforming surgical environment during minimally invasive surgery is important for intra-operative guidance. One key component of reliable reconstruction is accurate camera pose estimation, which is challenging for monocular cameras due to the paucity of reliable salient features, coupled with narrow baseline during surgical navigation. With recent advances in miniaturized MEMS sensors, the combination of inertial and vision sensing can provide increased robustness for camera pose estimation particularly for scenes involving tissue deformation. The aim of this work is to propose a robust framework for intra-operative free-form deformation recovery based on structure-from-motion. A novel adaptive Unscented Kalman Filter (UKF) parameterization scheme is proposed to fuse vision information with data from an Inertial Measurement Unit (IMU). The method is built on a compact scene representation scheme suitable for both surgical episode identification and instrument-tissue motion modelling. Detailed validation with both synthetic and phantom data is performed and results derived justify the potential clinical value of the technique.**

I. Introduction

In Minimally Invasive Surgery (MIS), accurate estimation of the motion of the endoscopic camera and reconstruction of its surrounding 3D anatomical structure are essential for intra-operative navigation and guidance. Dynamic motion stabilization and controlling the motion of surgical instruments by visual servoing are essential in robotically assisted laparoscopic surgery. To this end, reliable tissue deformation recovery is the prerequisite of all these techniques. Thus far, approaches based on stereo reconstruction, Simultaneous Localization and Mapping (SLAM) and Structure from Motion (SFM) [1] have been used but they are mainly applied to static scenes.

For the reconstruction of deforming surfaces observed with moving cameras, the assumption of periodic tissue motion has been made [2] [3] to recondition the problem such that it is solvable. However, this assumption is not always realistic for in-vivo MIS procedures where the tissue motion is too complex to be expressed by a single model. To overcome this problem, a free form deformation recovery approach using Gaussian Mixture Model based SFM [4] has been proposed without the use of explicit models on deformation.

In practice, the prerequisite for reliable deformation recovery is accurate camera pose estimation. Direct application of the commonly used vision techniques for pose estimation during MIS has significant problems due to the paucity of reliable salient features to track coupled with changing visual appearance of the surgical environment and what is often a narrow baseline during surgical navigation. With increasing miniaturization and reliability of Microelectromechanical Systems (MEMS) based inertial sensors and gyroscopes, their integration with normal surgical instruments is a reality. The integration of information from vision and inertial sensing can provide increased robustness for feature tracking and reduced ambiguity in camera pose estimation. This offers the possibility of developing more practical approaches for SFM based on the complementary nature of these two sensing modalities. More specifically, inertial sensors have large measurement uncertainty during slow motion and lower relative uncertainty at high velocities, whereas vision-based approaches are unable to cope with fast and unpredictable motions.

In recent years, the combination of inertial sensors with camera tracking has received increased attention. The monocular SLAM framework has been combined with IMU measurement data to enhance the robustness of the classical SLAM method [5] or estimate the scale parameter [6]. The Extended Kalman Filter (EKF) has been extensively used to combine visual and inertial information for robust egomotion [7], [8]. Multi-rate Kalman filters have also been designed to deal with data at different sampling rates for camera motion and structure estimation [9].

The use of the gravity reference provided either by inertial sensors or estimated based on the vanishing points from visual data, has been explored to reduce the minimum number of points required for camera motion estimation [10] [11]. In [12] the gravity was used as a vertical reference to estimate the camera focal distance based on the orthogonality between the vertical reference and the vanishing points of horizontal lines, using only one vanishing point.

A number of closed-form solutions [13] have been derived for the determination of attitude (roll and pitch angles), speed, absolute scale and bias. This is based on a system comprising of an IMU and a single monocular camera. These different contexts include the case of multiple or single features, the presence or absence of gravity and bias in inertial measurements.

The combination of inertial sensors with monocular cameras for real-time robust camera motion estimation and reconstruction over trajectories with challenging dynamics is presented in [14]. In [15] two autonomous Micro Air Vehicles (MAVs), each fitted with an IMU and a monocular camera are used to perform collaborative stereo. The IMU measurements are fused with the vision data to resolve

S. Giannarou, Z. Zhang and G.-Z. Yang are with the Hamlyn Centre for Robotic Surgery, Imperial College London, UK, E-mail:`stamatia.giannarou@imperial.ac.uk`, `z.zhang@imperial.ac.uk`, `g.z.yang@imperial.ac.uk`

978-1-4673-1736-8/12/S31.00 ©2012 IEEE

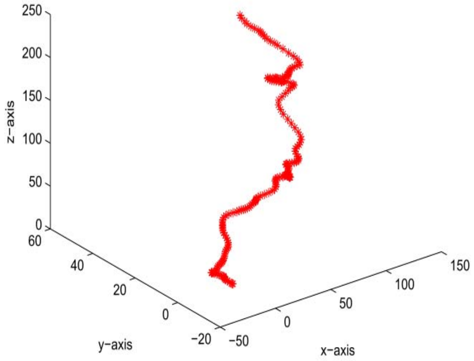


Fig. 1: Sample camera trajectory from the synthetic data.

scale ambiguity. A real-time hybrid solution to articulated 3D arm motion tracking for home-based rehabilitation by combining visual and inertial sensors is introduced in [16]. In robotic surgery, the tracking of a hand-held input devices is demonstrated in [17] where inertial data and and 2D locations of optical markers in the stereo camera images are fused by using an EKF.

The aim of this paper is to present a novel approach to robust 3D reconstruction of a deforming surgical scene observed with a projective monocular camera. A novel adaptive Unscented Kalman Filter (UKF) parameterization scheme is proposed to fuse vision information with data from an Inertial Measurement Unit (IMU). A simplified UKF process model has been designed to reduce the computational complexity, thus making it more amenable to real-time implementation. Interference from the accelerometer is adaptively compensated and the UKF also incorporates angular velocity measurements from the gyroscope of the IMU. The proposed deformation recovery framework is built on a compact scene representation scheme that is suitable for both surgical episode identification and instrument-tissue motion modelling. Detailed validation is provided on both synthetic and phantom data.

II. Methods

Prior to deformation recovery, temporal segmentation of the video sequence is applied to ensure coherent episodes are derived. This avoids tracking across episode boundaries, which are both technically difficult and practically meaningless. To this end, a succinct content-based data representation scheme that is suitable for both surgical episode identification and instrument-tissue motion modelling is used [18]. Each surgical episode is reconstructed separately and the motion characteristics of salient features are used to identify tissue deformation in response to instrument interaction.

A. Structure from Motion

For reliable and persistent feature tracking, an affine-invariant anisotropic region detector [19] is employed. An EKF parameterization scheme is used to adaptively adjust the optimal templates of the detected regions, enabling accurate identification and matching of a set of tracked features over a series of video frames [18].

Given a set of feature tracks W between the first and the last frame of a detected episode, the camera motion R and t is required for the estimation of the scene structure S. In this work, the 5-point algorithm [20] is applied in the above two views to generate a number of hypothesis for the essential matrix $E \equiv [t]_x R$ which are scored based on the reprojection error over all the points tracked. For subsequent episode frames where the camera baseline is not wide enough for triangulation, the method switches to the PnP method proposed in [21] for perspective pose estimation in order to derive the relative camera pose. For outlier removal, both the 5-point algorithm and the PnP method are used in conjunction with Random Sampling Consensus (RANSAC). Each camera pose and the estimated structure are refined by an iterative non-linear optimization step on the inlier subset. Common features tracked between consecutive episodes are used to resolve the relative scale between the estimated structures.

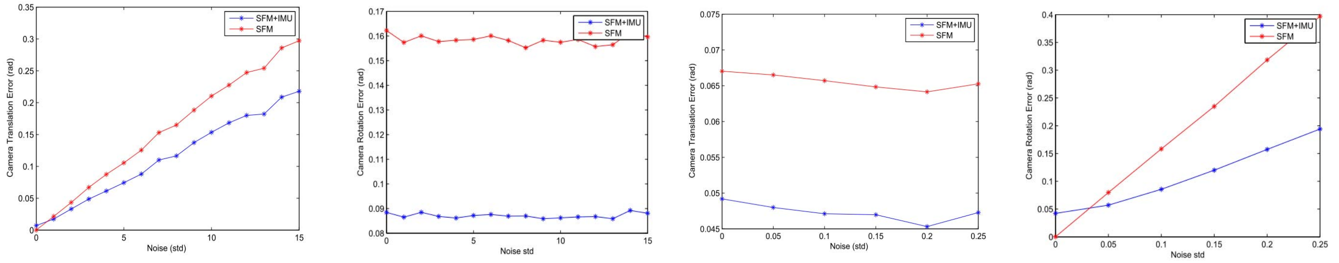
The above SFM framework is based on the assumption of a static scene. The aim of the proposed work is to simultaneously recover both camera motion and tissue deformation. To this end, features on deforming areas are automatically identified as outliers while the inliers correspond to static scene parts. The static areas are used to estimate the camera motion and are successfully reconstructed with the above framework. However, the accuracy in the deforming areas is low. In order to recover an accurate 3D structure of the entire observed environment, deformable areas are localized and their 3D shape is further refined [4]. To this end, inertial and vision measurements are fused by using the Unscented Kalman Filter (UKF).

B. Fusion of Visual and Inertial Measurement Data

For the proposed UKF scheme, the state of the system is composed of the position and orientation of the camera. Position is described with Cartesian positions, their velocities and accelerations. The orientation is represented with quaternion.

$$x_t = \begin{pmatrix} b_t \\ \dot{b}_t \\ \ddot{b}_t \\ q_t \end{pmatrix}^T \tag{1}$$

where b_t, $\dot{b}$, $\ddot{b}$ are the camera position, velocity and acceleration in the reference coordinate system, respectively. q_t is the orientation of the camera with respect to the reference coordinate system. The unit quaternion q_t is selected to represent the orientation, because it does not suffer from the singularity problem associated with Euler angles/rotation matrix. Moreover, the quaternion is computationally efficient as opposed to Euler angles, as it does not involve trigonometric functions to compute the rotation matrix. It has only one redundant parameter, as opposed to six in the rotation matrix. The relationship between x_t and x_{t-1} can be written



(a) (b) (c) (d)

Fig. 2: Validation results on synthetic data (a) Camera translation error and (b) rotation error with varying translation noise and constant rotation noise ($std = 0.1rad$) (c) Camera translation error and (d) rotation error with varying rotation noise and constant translation noise ($std = 3mm$).

as:

$$x_t = Fx_{t-1} + e_t \tag{2}$$

where

$$F = \left[\begin{array}{cccc} I_{3\times3} & I_{3\times3}\Delta t & I_{3\times3}\Delta t^2/2 & 0 \\ 0 & I_{3\times3} & I_{3\times3}\Delta t & 0 \\ 0 & 0 & I_{3\times3} & 0 \\ 0 & 0 & 0 & \Theta_t(\Delta t) \end{array} \right]$$

Here, $I_{3\times3}$ is the identity matrix of order 3, Δt is the sampling rate, and e_t is the process noise which is assumed to be zero mean Gaussian noise with covariance matrix Q. $\Theta_t(\Delta t) = \exp\{\frac{1}{2}\mathcal{R}(\omega^t)\Delta t\}$, where $\omega_t = \left(\omega_x^t,\ \omega_y^t,\ \omega_z^t\right)$ is the gyroscope measurement at time t resolved in the camera coordinate system and

$$\mathcal{R}(\omega_t) = \left[\begin{array}{cccc} 0 & -\omega_z^t & \omega_y^t & \omega_x^t \\ \omega_z^t & 0 & -\omega_x^t & \omega_y^t \\ -\omega_y^t & \omega_x^t & 0 & \omega_z^t \\ -\omega_x^t & -\omega_y^t & -\omega_z^t & 0 \end{array} \right]$$

We can rewrite the matrix exponential Θ_t using its Taylor series expansion as:

$$\begin{aligned} \Theta_t(\Delta t) =& I_{4\times4} + \frac{1}{2}\mathcal{R}(\omega_t)\Delta t + \frac{1}{2!}(\frac{1}{2}\mathcal{R}(\omega_t)\Delta t)^2 \\ &+ \frac{1}{3!}(\frac{1}{2}\mathcal{R}(\omega_t)\Delta t)^3 + \cdots \end{aligned} \tag{3}$$

where $I_{4\times4}$ is the identity matrix of dimension 4. The matrix $\mathcal{R}(\omega_t)$ has the following properties,

$$\begin{aligned} \mathcal{R}(\omega_t)^2 &= -|\omega_t|^2 \cdot I_{4\times4} \\ \mathcal{R}(\omega_t)^3 &= -|\omega_t|^2 \cdot \mathcal{R}(\omega_t) \\ \mathcal{R}(\omega_t)^4 &= |\omega_t|^4 \cdot I_{4\times4} \\ \mathcal{R}(\omega_t)^5 &= |\omega_t|^4 \cdot \mathcal{R}(\omega_t) \\ \mathcal{R}(\omega_t)^6 &= -|\omega_t|^6 \cdot I_{4\times4} \end{aligned} \tag{4}$$

By substituting these properties into (3), we can get

$$\begin{aligned} \Theta_t(\Delta t) =& \cos(\frac{|\omega_t|\Delta t}{2}) \cdot I_{4\times4} \\ &+ \frac{1}{|\omega_t|}\sin(\frac{|\omega_t|\Delta t}{2}) \cdot \mathcal{R}(\omega_t) \end{aligned} \tag{5}$$

The position $z_{b,t}^v$ and orientation $z_{q,t}^v$ measurements obtained from the SFM and the acceleration $z_{a,t}^{imu}$ and orientation $z_{q,t}^{imu}$ measured by the accelerometer and gyroscope measurement integration, respectively, are combined in the following measurement model:

$$z_t = \begin{pmatrix} z_{b,t}^v \\ z_{q,t}^v \\ z_{a,t}^{imu} \\ z_{q,t}^{imu} \end{pmatrix} = h(x_t) + v_t = h(x_t) + \begin{pmatrix} v_{b,t}^v \\ v_{q,t}^v \\ v_{a,t}^{imu} \\ v_{q,t}^{imu} \end{pmatrix} \tag{6}$$

where v_t is assumed to be zero mean Gaussian noise with covariance matrix V. The position and orientation measurement model is constructed as:

$$z_t = \begin{pmatrix} z_{b,t}^v \\ z_{q,t}^v \\ z_{q,t}^{imu} \end{pmatrix} = Hx_t + \begin{pmatrix} v_{b,t}^v \\ v_{q,t}^v \\ v_{q,t}^{imu} \end{pmatrix} \tag{7}$$

where

$$H = \left[\begin{array}{cccc} I_{3\times3} & 0_{3\times3} & 0_{3\times3} & 0_{4\times4} \\ 0_{3\times3} & 0_{3\times3} & 0_{3\times3} & I_{4\times4} \\ 0_{3\times3} & 0_{3\times3} & 0_{3\times3} & I_{4\times4} \end{array} \right]$$

The accelerometer signal $z_{a,t}^{imu}$ contains measurements of the body acceleration vector $\ddot{b}_t$ and the gravity vector g, both expressed in the sensor coordinate system as:

$$z_{a,t}^{imu} = q_t^{-1} \otimes \left(\ddot{b}_t + g\right) \otimes q_t + v_{a,t}, \tag{8}$$

where $\otimes$ is the quaternion multiplication.

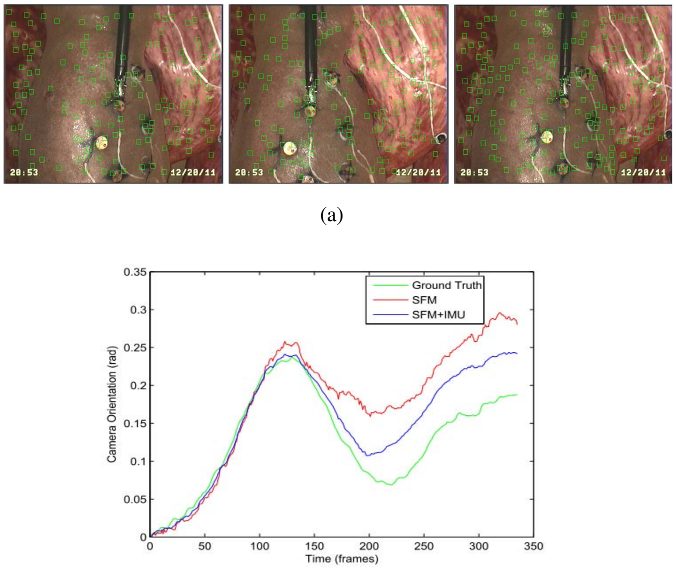
It is well-known that at slow motion, the MEMS accelerometer is unable to sense accurately the camera movement $\ddot{b}_t$, which is drowned in the noise $v_{a,t}^{imu}$. In that case, the acceleration $\ddot{b}_t$ should be ignored for the estimation of the measurement $z_{a,t}^{imu}$. In order to compensate the interference from the accelerometer an adaptive model is proposed for $z_{a,t}^{imu}$, defined as:

$$z_{a,t}^{imu} = \begin{cases} q_t^{-1} \otimes \left(\ddot{b}_t + g\right) \otimes q_t + v_{a,t}, & \text{if } \Big|\|g\| - \|z_{a,t}^{imu}\|\Big| > \epsilon_A \\ q_t^{-1} \otimes g \otimes q_t + v_{a,t}, & \text{otherwise} \end{cases} \tag{9}$$

where ϵ_A was set to $0.1g$ in our work.

C. Deformation Localization and Recovery

For deformation localization and recovery, probabilistic motion modelling is used to represent the motion of the tracked features. Rather than using explicit priors for motion modelling, only a weak constraint of locally similar motion is assumed. This makes the technique more generalizable to *in vivo* cases with unknown deformation. The motion of each tracked feature is modelled as a mixture of Gaussian distributions and the motion models of the features are clustered to identify areas of coherent motion within the episode [18]. The static part of the observed scene corresponds to the cluster that includes the highest number of inliers extracted from the SFM framework. The remaining clusters that have survived the refinement process represent independently deforming areas. The initial structure estimated from the SFM framework explained above is refined for each independently moving area, individually. The only assumption of the proposed approach to recover free-form deformation is that prior to the tool-tissue interaction, the camera navigates in the surgical environment in order to estimate the initial 3D structure of the scene while it is static [4].



(a)

(b)

Fig. 3: (a) Episode border frames from Sequence 1 of the liver phantom data with tracked points represented by green squares (b) Camera orientation estimation.

III. Results

In order to assess the practical value of the proposed framework, quantitative evaluation has been performed on synthetic and phantom data. Related work is not suitable for comparison here as assumptions of static background [1], known tissue motion model [2] or priors on the camera pose and surface shape are made in previous studies.

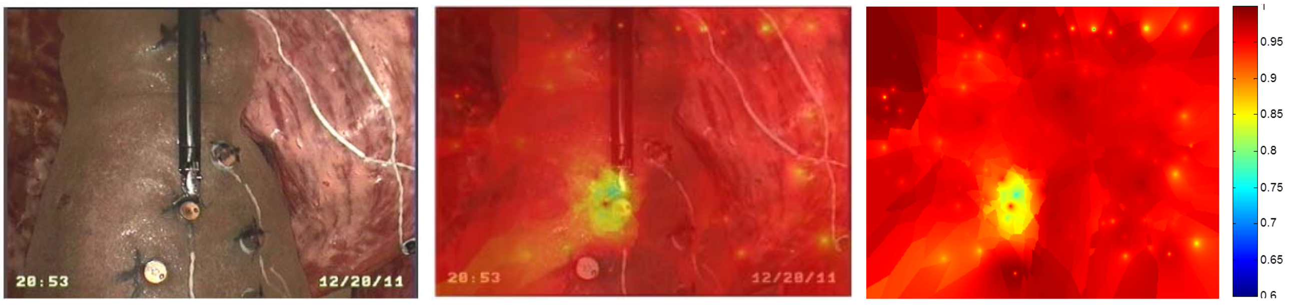
For the synthetic data, a set of camera trajectories was generated using an optical tracking device (Northern Digital Inc., Ontario, Canada). To obtain the position of the camera, a rigid stereo laparoscope fitted with eight optical markers was used. The position and orientation of the centre of the camera relative to the optical markers were acquired using standard hand-eye calibration. IMU measurement data were collected by attaching a Xsens MEMS-based miniature IMU on the rigid body. A sample camera trajectory is shown in Fig. 1.

In order to evaluate the robustness of the method to noise, the generated camera trajectories were contaminated with noise and used as input to our adaptive fusion framework. Observation error was added to the camera position in the x, y and z-axes, in the form of Gaussian noise with zero mean and standard deviation ranging from 0 to 15 mm. In a similar way, Gaussian noise with zero mean and standard deviation ranging from 0 to 0.25 rad was added to the camera orientation with respect to the x, y and z-axis. For each noise level, we run 50 trials and the final result is the mean error from all the trials. The camera rotational error is estimated as the smallest angle of rotation that can bring the estimate to the true value. The translational error is the deviation of the estimated translation direction from the true value. The ability of the proposed framework to suppress the noise in the estimation of the camera pose is illustrated in Fig.2, where the error is always lower when visual data is combined with inertial measurement data compared to when only visual information is used. The error of the proposed adaptive fusion method at zero noise is slightly higher than zero due to measurement noise from the UKF which makes the fusion result more blurred than the vision.

The performance of the proposed framework was further evaluated on a liver phantom made of silicon rubber. A laparoscope was used to capture two sequences of video data with the camera navigating around the phantom in the

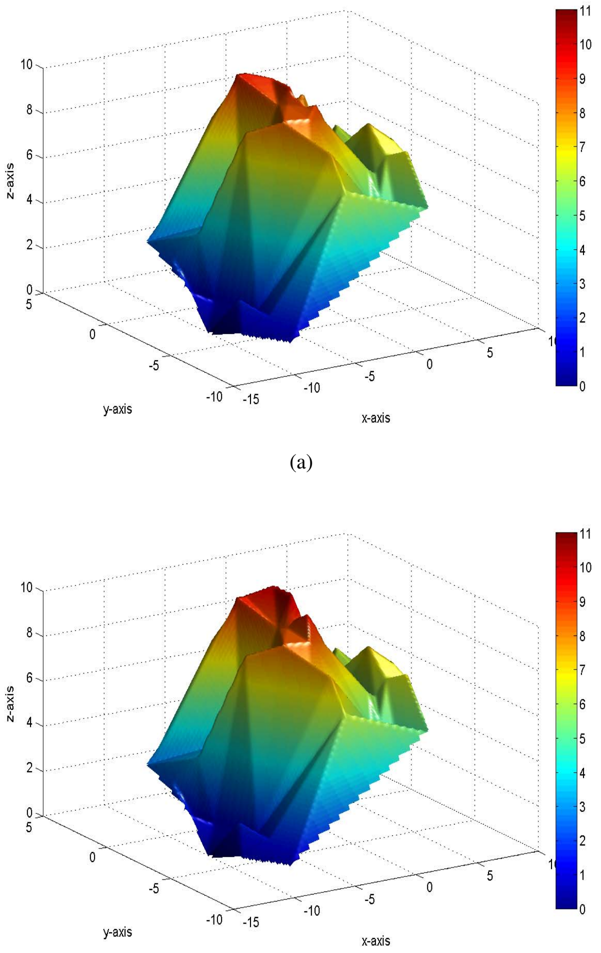
TABLE I: Validation results on phantom data

	Deform. Recovery Error (mm)		**Surface Reconstr. Error (mm)**		**Rotational Error (rad)**		**Translational Error (rad)**	
Method	*SFM*	*SFM+IMU*	*SFM*	*SFM+IMU*	*SFM*	*SFM+IMU*	*SFM*	*SFM+IMU*
Seq. 1	5.26	4.70	3.6	3.5	0.051	0.026	0.31	0.30
Seq. 2	5.10	4.90	5.89	4.85	0.013	0.003	0.49	0.48



(a) (b) (c)

Fig. 4: (a) Sample frame from Sequence 2 of the liver phantom data (b) Deformation localization (c) Clustering of the motion models in the scene.

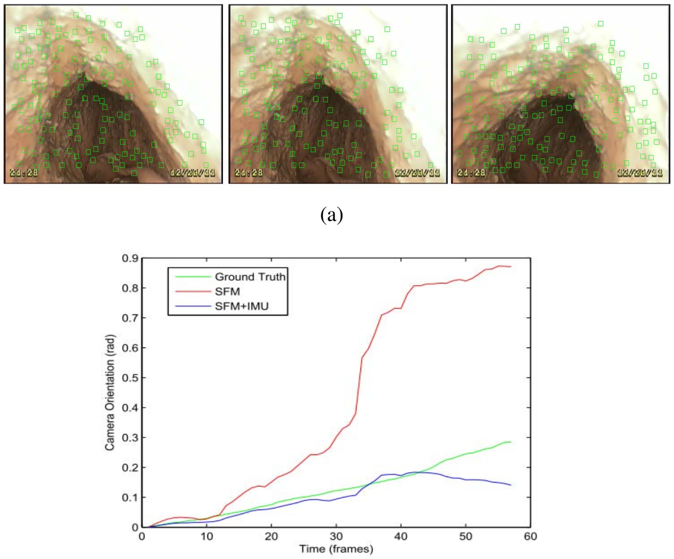


(a)

(b)

Fig. 5: (a) Scene structure prior to the tool-phantom interaction (b) Refinement of structure (a) after deformation recovery. Deformation corresponds to the dark red area at the peak of the surface.

presence of tool-tissue interaction. Validation was performed by measuring the accuracy in the estimation of the deformation recovery and the 3D surface reconstruction in the camera space and the error in the estimated camera motion as the laparoscope navigated around the phantom. Ground truth data of the camera pose was collected using an optical



(a)

(b)

Fig. 6: (a) Sample frames from the colon phantom data with tracked points represented by green squares (b) Camera orientation estimation.

tracking device in the same way as explained above. To obtain the ground truth of the 3D structure, optical markers were attached to the phantom surface. For each optical marker, the salient feature on the phantom surface closest to the marker was identified. The 3D position of these features estimated with the proposed framework was compared to the ground truth 3D structure to estimate the 3D surface reconstruction error. The ground truth of the deformation of the phantom surface was obtained by estimating the displacement in the 3D position of the markers closer to the point where the tool-phantom interaction took place.

Episode border frames from Sequence 1 of the liver data and tracked affine-invariant anisotropic regions are illustrated in Fig. 3(a). A set of 150 regions were detected in the first frame of each episode and those that have been successfully tracked during each episode are used for camera pose estimation and reconstruction. The camera orientation curves in Fig. 3(b) show that the proposed adaptive fusion framework outperforms the SFM approach and gives a camera pose estimation close to the ground truth.

Areas of coherent motion within an episode with tool-phantom interaction from Sequence 2 of the liver data, are graphically classified using the colormap in Fig. 4(c) to demonstrate the similarity between the motion of scene points and a reference point (the upper left corner of the scene). In Fig. 4(b) the similarity colormap is superimposed on the episode frame in Fig. 4(a) to illustrate the deformation localization result. The reconstructed surfaces when the scene is static and when the maximum deformation is applied are presented in 5(a) and 5(b), respectively. The surface of the phantom liver is generated by interpolating the 3D position of the tracked salient features. The colormap in Fig. 4 represents the distance of the surface from the camera. Due to tool-phantom interaction the distance between the points on the deformed area and the camera has increased and corresponds to the dark red area at the peak of the surface in Fig. 4(b).

Table 1 presents the deformation recovery error when the maximum deformation is applied, for each phantom sequence. The surface reconstruction and camera motion errors are the mean errors for the whole sequence. The performance improvement gained by the proposed adaptive fusion framework is evident. The high error values in Table 1 are justified by the narrow baseline between the reconstructed frames as it can be noticed in Fig. 3(a).

The robustness of the proposed framework in a more challenging navigation environment such as the colon phantom in Fig. 6(a) is shown in Fig. 6(b). In the above scenario, the paucity of reliable features and the narrow baseline between the borders of the detected episodes makes the camera pose estimation difficult. Fig. 6(b) shows that the fusion of the vision and inertial data reduces significantly the error in the estimation of the camera orientation.

IV. Conclusions

In this paper, we have proposed a novel approach for fusion of vision and inertial measurement data to facilitate robust recovery of free-form deformation of the surgical environment in MIS. This represents one of the first attempts to combine vision and inertial sensing for robust pose estimation in MIS. Unlike previous approaches, the proposed framework does not impose explicit constraints on tissue deformation, allowing realistic free-form deformation recovery. The proposed framework has been tailored for adaptive motion stabilization and visual servoing in robotically assisted laparoscopic surgery. It can also be used to maintain consistent force of imaging probes such as point based confocal laser microscopy on the tissue surface to prevent distortion to the image morphology due to excessive probe pressure. Furthermore, the proposed method could help the diagnosis of diseases such as liver cirrhosis by measuring the modulus of stiffness of the liver during instrument-tissue interaction. Results derived from validation on synthetic and phantom data demonstrate the intrinsic accuracy achievable and the potential clinical value of the technique.

References

[1] D. Mirota, H. Wang, R. Taylor, M. Ishii, and G. Hager, "Toward video-based navigation for endoscopic endonasal skull base surgery," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 5761, 2009, pp. 91–99.

[2] M. Hu, G. Penney, D. Rueckert, P. Edwards, R. Bello, and R. C. et al, "Non-rigid reconstruction of the beating heart surface for minimally invasive cardiac surgery," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 1, 2009, pp. 34–42.

[3] P. Mountney and G.-Z. Yang, "Motion compensated slam for image guided surgery," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 2, 2010, pp. 496–504.

[4] S. Giannarou and G.-Z. Yang, "Tissue deformation recovery with gaussian mixture model based structure from motion," in *Workshop on Augmented Environments for Computer-Assisted Interventions (AE-CAI), International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2011.

[5] F. Servant, P. Houlier, and E. Marchand, "Improving monocular plane-based slam with inertial measures," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 3810–3815.

[6] G. Nützi, S. Weiss, D. Scaramuzza, and R. Siegwart, "Fusion of imu and vision for absolute scale estimation in monocular slam," *Journal of Intelligent and Robotic Systems*, vol. 61, pp. 287–299, 2011.

[7] R. Voigt, J. Nikolic, C. Hurzeler, S. Weiss, L. Kneip, and R. Siegwart, "Robust embedded egomotion estimation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011, pp. 2694–2699.

[8] J.-P. Tardif, M. George, M. Laverne, A. Kelly, and A. Stentz, "A new approach to vision-aided inertial navigation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 4161–4168.

[9] P. Gemeiner, P. Einramhof, and M. Vincze, "Simultaneous motion and structure estimation by fusion of inertial and vision data," *The International Journal of Robotics Research*, vol. 26, no. 6, pp. 591–605, 2007.

[10] F. Fraundorfer, P. Tanskanen, and M. Pollefeys, "A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 269–282.

[11] O. Naroditsky, X. Zhou, J. Gallier, S. Roumeliotis, and K. Daniilidis, "Two efficient solutions for visual odometry using directional correspondence," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 4, pp. 818–824, 2011.

[12] J. Lobo and J. Dias, "Vision and inertial sensor cooperation using gravity as a vertical reference," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 12, pp. 1597–1608, 2003.

[13] A. Martinelli, "Vision and imu data fusion: Closed-form solutions for attitude, speed, absolute scale and bias determination," *IEEE Transactions on Robotics*, no. 99, pp. 1–17, 2011.

[14] L. Kneip, M. Chli, and R. Siegwart, "Robust real-time visual odometry with a single camera and an imu," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2011, pp. 16.1–16.11.

[15] M. W. Achtelik, S. Weiss, M. Chli, F. Dellaert, and R. Siegwart, "Collaborative stereo," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, sept. 2011, pp. 2242–2248.

[16] Y. Tao, H. Hu, and H. Zhou, "Integration of vision and inertial sensors for 3d arm motion tracking in home-based rehabilitation," *The International Journal of Robotics Research*, vol. 26, no. 6, pp. 607–624, 2007.

[17] A. Tobergte, M. Pomarlan, G. Passig, and G. Hirzinger, "An approach to ulta-tightly coupled data fusion for handheld input devices in robotic surgery," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 2424–2430.

[18] S. Giannarou and G.-Z. Yang, "Content-based surgical workflow representation using probabilistic motion modeling," in *International Workshop on Medical Imaging and Augmented Reality*, 2010, pp. 314–323.

[19] S. Giannarou, M. Visentini-Scarzanella, and G.-Z. Yang, "Affine-invariant anisotropic detector for soft tissue tracking in minimally invasive surgery," in *IEEE International Symposium on Biomedical Imaging*, 2009, pp. 1059–1062.

[20] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 6, pp. 756–770, June 2004.

[21] F. Moreno-Noguer, V. Lepetit, and P. Fua, "Accurate non-iterative o(n) solution to the pnp problem," in *IEEE International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.