Simultaneous Stereoscope Localization and Soft-Tissue Mapping for Minimal Invasive Surgery

Peter Mountney[1], Danail Stoyanov[1], Andrew Davison[1], and Guang-Zhong Yang[1,2]

[1] Royal Society/Wolfson Foundation Medical Image Computing Laboratory,
[2] Department of Surgical Oncology and Technology
Imperial College, London SW7 2BZ, UK
{peter.mountney, danail.stoyanov, andrew.davison,
g.z.yang}@imperial.ac.uk
http://vip.doc.ic.ac.uk

Abstract. Minimally Invasive Surgery (MIS) has recognized benefits of reduced patient trauma and recovery time. In practice, MIS procedures present a number of challenges due to the loss of 3D vision and the narrow field-of-view provided by the camera. The restricted vision can make navigation and localization within the human body a challenging task. This paper presents a robust technique for building a repeatable long term 3D map of the scene whilst recovering the camera movement based on Simultaneous Localization and Mapping (SLAM). A sequential vision only approach is adopted which provides 6 DOF camera movement that exploits the available textured surfaces and reduces reliance on strong planar structures required for range finders. The method has been validated with a simulated data set using real MIS textures, as well as *in vivo* MIS video sequences. The results indicate the strength of the proposed algorithm under the complex reflectance properties of the scene, and the potential for real-time application for integrating with the existing MIS hardware.

1 Introduction

In surgery, the increasing use of MIS is motivated by the benefit of improved therapeutic outcome combined with reduced patient trauma and hospitalization. The technique is increasingly being used to perform procedures that are otherwise prohibited by the confines of the operating environment. MIS also offers a unique opportunity for deploying sophisticated surgical tools that can greatly enhance the manual dexterities of the operating surgeon. Despite the benefit of MIS in terms of patient recovery and surgical outcome, the practical deployment of the technique is complicated by the complexity of instrument control and difficult hand-eye coordination. Due to the large magnification factors required for performing MIS tasks, the field-of-view of the laparoscope cameras is usually very limited. This results in restricted vision which can affect the visual-spatial orientation of the surgeon and the awareness of the peripheral sites.

In order to facilitate the global orientation of the target site, a number of spatial localization techniques have been developed. These include the use of pre-operative imaging combined with 2D/3D registration such that the underlying structure and morphology of the soft-tissue can be provided. To cater for tissue deformation,

R. Larsen, M. Nielsen, and J. Sporring (Eds.): MICCAI 2006, LNCS 4190, pp. 347 – 354, 2006.

structure from light [1] or motion sensors such as mechanically or optically based accelerometers [2, 3] are used. With the increasing availabilities of stereo-laparoscope cameras, detailed 3D motion and structure recovery techniques based on stereo vision have also be proposed recently [4, 5]. The major advantage of the optical methods is that they do not require additional modifications to the existing MIS hardware, and thus are easily generalizable to different clinical settings. One of the limitations of the above techniques is that they only consider information captured in the current field-of-view. Global information that is implicitly captured by the moving laparoscope camera is typically discarded. An exception to this is [6], where a map is built containing global information. The camera estimation is based on structure from motion, which is susceptible to drift.

The purpose of this study is to investigate the use of SLAM for simultaneous stereoscope localization and soft tissue mapping. In essence, the SLAM problem is concerned with the estimation of moving sensor while building a reconstruction of what it observes. The advantage of the method is that it builds a long-term map of the features with minimal drift, allowing localization of the sensor after long periods of feature neglect [7, 8]. This is particularly useful for laparoscope with restricted vision in that a global map of the operating field-of-view can be integrated with moving stereo vision. In this study, a sequential vision only approach is adopted which provides 6 DOF camera movement that exploits the available textured surfaces and reduced reliance on strong planar structures required for range finders. More importantly, it provides the potential for real-time application for integrating with the existing MIS hardware.

2 Methods

2.1 Building a Statistical Map

For stereoscope localization and soft tissue mapping, our aim is to recover the trajectory of the stereoscope and build a map of the environment. In a Kalman filter framework, the overall state of the system x is represented as a vector. This vector is partitioned into the state $\hat{x}_v$ of the camera and the states $\hat{y}_i$ of the features which make up the map. Crucially, the state vector is accompanied by a single covariance matrix which can also be partitioned as follows:

$$\hat{x} = \begin{pmatrix} \hat{x}_v \\ \hat{y}_1 \\ \hat{y}_2 \\ \vdots \end{pmatrix}, \qquad P = \begin{bmatrix} P_{xx} & P_{xy_1} & P_{xy_2} & \cdots \\ P_{y_1x} & P_{y_1y_1} & P_{y_1y_2} & \cdots \\ P_{y_2x} & P_{y_2y_1} & P_{y_2y_2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{1}$$

The role of the covariance matrix is to represent the uncertainty to first order in all the quantities in the state vector. Feature estimates $\hat{y}_i$ can be freely added to or deleted from the map as required causing x and P to grow or shrink dynamically. In this framework, x and P are updated in two steps: 1) the prediction step uses a motion model to calculate how the camera moves during surgery and how its position

uncertainty increases; 2) the measurement step describes how the map and camera position uncertainties can be reduced when new input from the stereoscope is processed. Maintaining a full feature covariance matrix P allows the camera to re-visit and recognize known areas after periods of neglect (this has been irrefutably proven in SLAM research).

With the proposed approach, camera calibration is required to estimate 3D positions from stereo images and feature locations in the image plane from 3D positions. Calibration is done assuming a pinhole camera model and using a closed form solution [9]. The centre of the camera rig is taken to be the left camera and the extrinsic parameters describe the translation and rotation of the right camera relative to the left camera. In MIS, the stereoscopic laparoscope is pre-calibrated before the surgical procedure and remains unchanged.

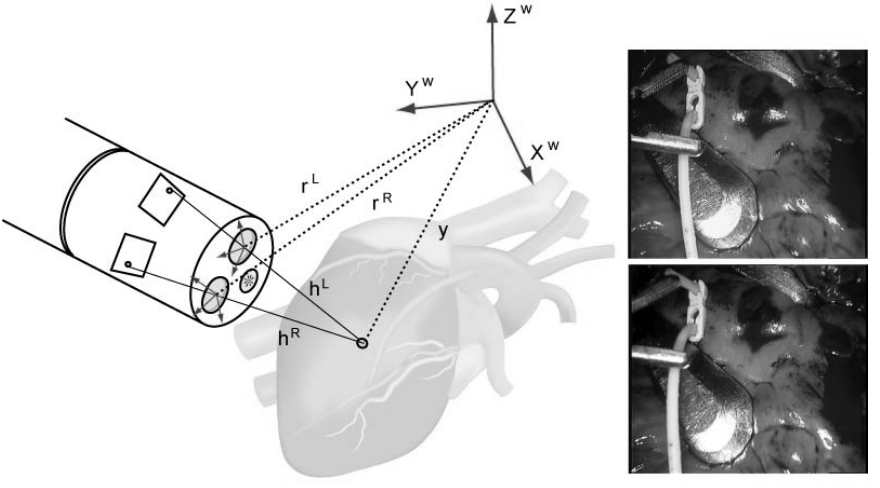


Fig. 1. Stereo-laparoscope camera geometry and an example image from a MIS scene. The figure illustrates the geometry between a global coordinate system, the local camera coordinates and a selected point from the map.

For the stereo-laparoscope camera, three coordinate frames illustrated in Fig. 1 are defined; W, fixed in the world, L, fixed with respect to the left camera and R, fixed with respect to the right camera.

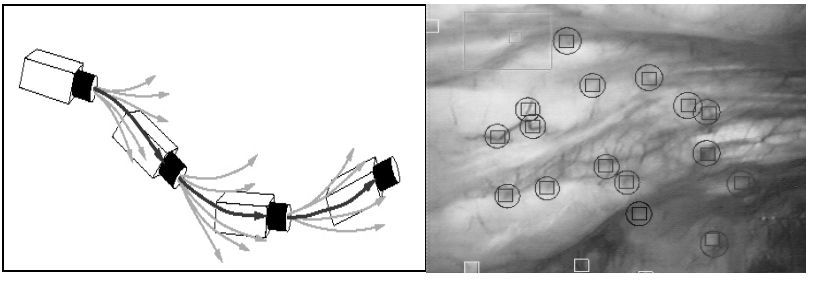
$$x_v = \begin{pmatrix} r^W \\ q^{WL} \\ v^W \\ \omega^W \end{pmatrix}, \qquad y_i = \begin{pmatrix} x_i^W \\ y_i^W \\ z_i^W \end{pmatrix} \tag{2}$$

We refer to x_v as the **state** of the camera rig. The state is made up of four parts; r^W a the position of the camera in the world coordinate system, q^{WL} the rotation of the camera in the world coordinate system, v^w is the linear velocity and ω^w the angular velocity. y_i refers to a feature consisting a 3D vector in XYZ Euclidean space.

2.2 Motion Model

In the case of a stereoscope moving during a MIS procedure, the motion model must take into account the unknown intentions of the operator. This unknown element can be modeled statistically by using a two part motion model. The first part is a deterministic element known as a "constant velocity, constant angular velocity model". This, however, does not mean that we assume that the camera moves at a constant velocity over all time. It only imposes that on average we expect its velocity and angular velocity to remain the same. The second part is stochastic and models the uncertainty in the surgeon's movement of the stereoscope. The uncertainty in the system is the acceleration modeled with a Gaussian profile. The implication of this model is that smoothness is implicitly imposed on the camera motion, very large accelerations are therefore relatively unlikely.

The rate of growth of uncertainty in this motion model is determined by the size of parameter P_n, and setting this to small or large values defines the smoothness of the motion we expect. With small P_n, we expect a very smooth motion with small accelerations, and would be well placed to track motions of this type but unable to cope with sudden rapid movements or changes in direction. High P_n means that the uncertainty in the system increases significantly at each time step, and while this gives the ability to cope with rapid accelerations the very large uncertainty means that a lot of good measurements must be made at each time step to constrain the estimates.



(a) **(b)**

Fig. 2. (a) Visualization of the model for 'smooth' motion: at each camera position a most likely path is predicted together with alternatives with small deviations. **(b)** A MIS scene where 25×25 pixel box represent feature patches detected using the Shi and Thomasi operator [10] and ellipses represent the estimated search regions for the landmarks.

2.3 Visual Feature

The *in vivo* anatomical structure is generally curved, thus making feature extraction more challenging than in man made environments. In [5], MSER features and weak gradient features were combined to create a dense 3D map of the heart. Features are tracked from frame to frame using a Lucas-Kanade tracker to recover the motion of the heart. These transient features work well for frame to frame tracking. However, in order to build a sustainable map, we require long term landmarks which are repeatable. A long term repeatable feature is one that is strongly salient and uniquely

identifiable. Previous work by Davison [6] has demonstrated long term features within a structured environment with a good degree of view point independence. This approach is used in this study to detect features using the saliency operator of Shi and Tomasi [9]. The feature is represented by a 25×25 pixel patch, and a normalized sum-of-squared difference correlation is used to match the feature in subsequent images. Specularities are removed through thresholding.

In the proposed framework, we aim to keep the number of visible features at a predetermined threshold to reduce reliance on weak features. A feature is "visible" if it is predicted to be in the current image. Features are added to the map if the number visible is less than this threshold. New features are detected in the left image and matched using normalized sum-of-squared difference in the right stereo image. Initialization is managed to prevent the same feature being tracked twice. Epipolar geometry is used to estimate the 3D position of the feature and all features are initialized with uniform uncertainty represented as a 3D Gaussian.

2.5 Measurement Model

Another important element of the proposed localization model is the measurement. The measurement model is the process for comparing the predicted SLAM map with the input from the stereoscope. The estimates x_p of camera position and y_i of feature position in 3D, allowing the position of the features to be predicted in the image plane. The position of a 3D feature relative to the camera is expected to be:

$$h_L^L = R^{LW}(y_i^W - r_L^W) \tag{3}$$

where R^{LW} is the rotation matrix transforming between the left camera frame L and world frame W. This is used to calculate (u_L, v_L) the predicted positions of the features in the left stereo image. The actual positions of the features in the images are obtained by actively searching the area around the predicted position. The search area is derived from the uncertainty of the feature's predicted position which is a 2D Gaussian p.d.f. over the image coordinates. Gating at three standard deviations provides an elliptic search window around the feature's predicted position.

3 Experimental Design

To validate the proposed method, a simulation with a virtual stereo camera moving through a texture mapped 3D world was rendered. The simulator as shown in Fig. 3 provides the ground truth data of known camera movement within a known environment. This allows the accuracy of the camera localization and mapping to be evaluated.

The camera motion was controlled so that the resultant inter-frame pixel motion did not exceed 20 pixels, which was consistent with observations from *in vivo* data. The virtual stereo camera rig was set up to replicate a stereo-laparoscope by taking similar camera intrinsic and extrinsic properties, notably the baseline was set to only 5mm. The environment contains a plane, which has been textured with an image taken from a robotic assisted totally endoscopic coronary artery bypass graft surgery

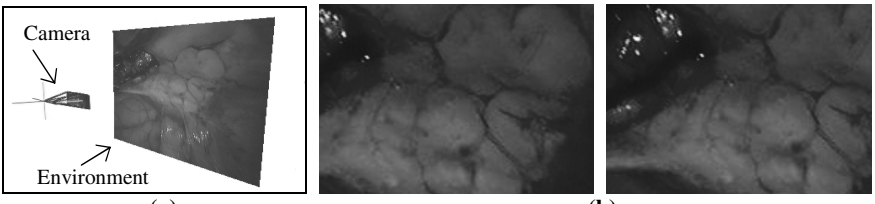


Fig. 3. An illustration of the simulation environment used to generate a stereo-laparoscopic video with known ground truth data for camera motion. A 3D rendition of the virtual world is shown in **(a)** and an example stereo pair taken from the virtual cameras is shown in **(b).**

to provide realistic image rendition. The use of a single planar model is not restrictive or degenerate as the proposed methods can be applied to more complex models.

In addition to synthetic simulations, the proposed technique has also been applied to real MIS videos. Since the ground truth data for the *in vivo* data is not available, qualitative analysis by forward tracking the motion and then reversing the video sequence is used to assess the internal consistency of the algorithm.

4 Results

In Fig. 4, the results of using the proposed technique to estimate the movement of the stereo-laparoscope over 176 frames of simulated video are provided. The stereo-laparoscope was moved by 1.5cm, 2cm and 0cm along the X, Y and Z axis respectively. Since no prior knowledge of the environment is taken, the initial estimations of feature positions have a large uncertainty. The uncertainty reduces as the stereoscope moves but creates a lag in the estimated movement. This is evident in the movement along the *X* axis. Small movement of around 1mm in the *Z* axis is a result of the narrow baseline of the stereoscope.

A potential problem with using a constant velocity motion models is the issue of dealing with sudden changes in direction. However, the results show the algorithm is robust to changes in direction. It can be shown that 87.7% of the recovered movement lies within three standard deviations of the ground truth, this represents a confidence interval of 99%.

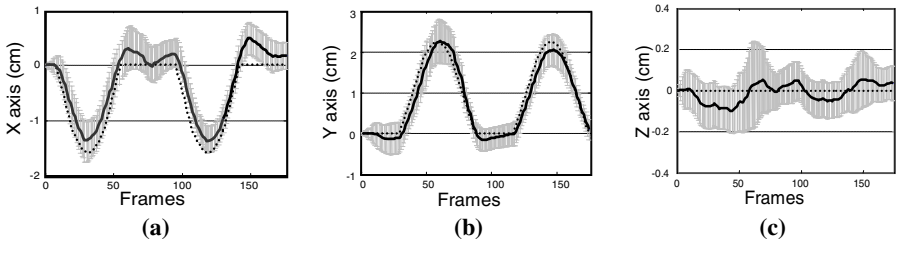


Fig. 4. Simulation based analysis of camera motion estimation. The graphs shown in **(a-c)** illustrate the recovered stereoscope movement in the *X*, *Y* and *Z* axes, respectively against the ground truth. The solid black line shows the estimated motion with the grey bars indicating the uncertainty associated with the estimate. The dotted line displays the ground truth motion.

For *in vivo* analysis, Fig. 5 shows example images from the left and right channels of the stereoscope, whereas Fig. 6 illustrates the recovered trajectory paths of the camera along the *X*, *Y* and *Z* axes in the world coordinate system. The original video footage is 79 frames and the reversed video is 79 frames long. It is evident from the graphs that the camera tracking closes the loop by returning the device close to its starting position. Finally, the SLAM map acquired from the *in vivo* sequence is shown in Fig. 7, along with an example image with selected features with their corresponding uncertainties. The appearance of features alters as the stereoscope and light source move. Feature matching is made more robust by using active search with the use of normalized sum-of-squared difference correlation to reduce data association errors.

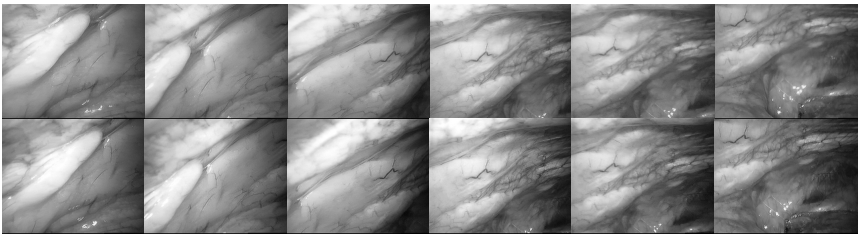


Fig. 5. Left (top) and right (bottom) stereo images taken from an *in vivo* stereo-laparoscope sequence that involves a change of camera viewing position and orientation

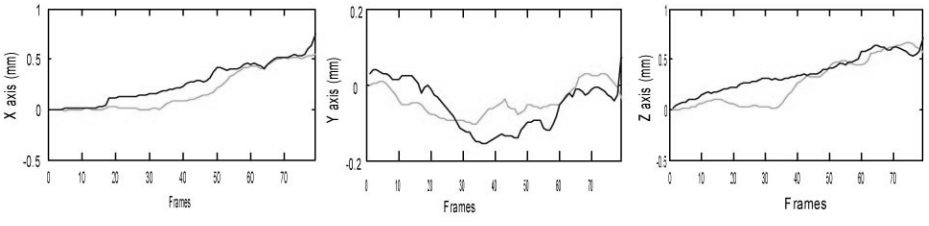


Fig. 6. *In vivo* analysis of the proposed techniques where the graphs show the recovered stereoscope movement along the *X*, *Y* and *Z* axes. Light grey lines represent the recovered motion in the forward sequence whereas the dark grey lines illustrate the recovered motion in the reverse direction.

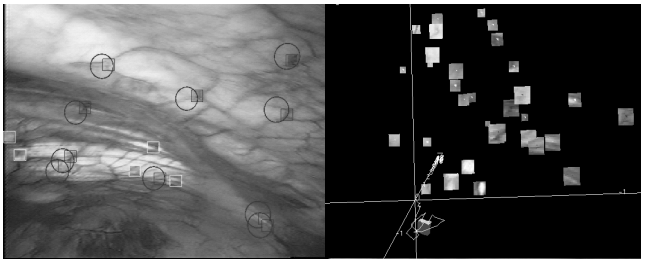


Fig. 7. Typical features selected in the left stereo image plane and the corresponding landmarks projected onto 3D coordinate system by using information built into the SLAM map

5 Discussion and Conclusions

In this paper, we have developed a technique to estimate the movement of the stereo-laparoscope during MIS and build a map of the anatomical structure. The method has been validated with a simulated data set using real MIS textures, as well as *in vivo* MIS video sequences. The results indicate the strength of the proposed algorithm under the complex reflectance properties of the scene. Accuracy can be further improved by incorporating information from the remaining stereo image into the measurement model, and directly cater for tissue deformation in the SLAM paradigm.

References

1. Keller K, Ackerman, J. Real-Time Structured Light Depth Extraction. in Proc of Three Dimensional Image Capture and Applications III SPIE, 2000, 11-18.
2. Hoffman J, Spranger M, Gohring D, Jungel M. Making Use of What You Don't See: Negative Information in Markov Localization. in Proc of Intelligent Robots and Systems, 2005.
3. Thrakal A WJ, Tomlin D, Seth N, Thakor N. . Surgical Motion Adaptive Robotic Technology (Smart): Taking the Motion out of Physiological Motion. in Proc of MICCAI, 2001, 317-325.
4. Chatila R, Laumond J. Position Referencing and Consistent World Modeling for Mobile Robots. in Proc of Robotics and Automation. 1985, 138-145.
5. Stoyanov D, Darzi, A., Yang, G-.Z. Dense 3d Depth Recovery for Soft Tissue Deformation During Robotically Assisted Laparoscopic Surgery. in Proc of MICCAI, 2004, 41-48.
6. Burschka D, Li M, Ishii M, Taylor RH, Hager GD. Scale-Invariant Registration of Monocular Endoscopic Images to Ct-Scans for Sinus Surgery. *Medical Image Analysis*, 2005, 9(5):413-439.
7. Davison AJ. Real-Time Simultaneous Localisation and Mapping with a Single Camera. in Proc of 9th IEEE ICCV, 2003, 1403.
8. Se S, Jasiobedzki, P. Instant Scene Modeler for Crime Scene Reconstruction. *IEEE A3DISS* 2005.
9. Zhang P, Milios EE, Gu J. Vision Data Registration for Robot Self-Localization in 3d. in Proc of Intelligent Robots and Systems, 2005, 2315-2320.
10. Shi J, Tomasi, C. Good Features to Track. in Proc of CVPR, 1994, 593-600.