Visual Odometry in Stereo Endoscopy by using PEaRL to handle Partial Scene Deformation[*]

Miguel Lourenço[1], Danail Stoyanov[2], and João P. Barreto[1,3]

[1] Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal
{miguel,jpbar}@isr.uc.pt
[2] Centre for Medical Image Computing, University College of London, London, UK
danail.stoyanov@ucl.ac.uk
[3] Perceive 3D, Coimbra, Portugal

Abstract. Stereoscopic laparoscopy provides the surgeon with the depth perception at the surgical site to facilitate fine micro-manipulation of soft-tissues. The technology also enables computer-assisted laparoscopy where patient specific models can be overlaid onto laparoscopic video in real-time to provide image guidance. To maintain graphical overlay alignment of image-guides it is essential to recover the camera motion and scene geometry during the procedure. This can be performed using the image data itself, however, despite of the mature state of structure-from-motion techniques, their application in minimally invasive surgery remains a challenging problem due non-rigid scene deformation. In this paper, we propose a method for recovering the camera motion of stereo endoscopes through a multi-model fitting approach which segments rigid and non-rigid structures at the surgical site. The method jointly optimizes the segmentation of image and uses the rigid structure to robustly estimate the motion of the laparoscope. Synthetic and *in-vivo* experiments show that the proposed algorithm outperforms RANSAC-based stereo visual odometry in non-rigid laparoscopic surgery scenes.

1 Introduction

Stereo laparoscopes are becoming increasingly popular in Minimally Invasise Surgey (MIS). The main reason behind their wide adoption is the possibility of recovering the 3D structure of the surgical site to provide the surgeon with depth perception of the operating field. Despite of being a difficult problem due to the dynamics of the medical environment that combine occlusions from the surgical instruments with strong specularities, several authors have already proposed efficient solutions for real-time computation of depth maps in medical endoscopy [1–3]. The obtained 3D structure can be used to align multimodal

[*] Miguel Lourenço and João Barreto want to thank *QREN-Mais Centro* by funding through the project *Novas Tecnologias para apoio à Saúde e Qualidade de Vida, Projecto A- Cirurgia e Diagnóstico Assistido por Computador Usando Imagem* and the Portuguese Science Foundation by funding through grant SFRH/BD/63118/2009.

information [4] within a global reference 3D coordinate system [5] and enhance robotic instrument control.

An early work on structure-from-motion (SfM) in laparoscopic surgery was developed by Burschka *et al.* [5] where a rigid environment was assumed due to the confines of the sinus in order to compute a 3D scene map for registration with pre-operative Computed Tomography (CT) patient models. For procedures targeting soft-tissue anatomies non-rigidity due to cardiac, respiratory or peristaltic motions can make such SfM impossible. Deformable SfM (DSfM) [3], motion compensated SLAM [6] and more recently Non-Rigid SfM [7] have been proposed for overcoming this problem but an inspection phase to build a rigid template of the scene and strong priors deformation are not always feasible. For example motion and anatomical deformation due to instrument interactions cannot be reliably modelled prior to surgery and significant practical challenges remain for robust SfM in MIS. It is also possible to incorporate position sensors for additional constraints but this involves difficult integration solutions [8]. Close work to ours was proposed by Roussos *et al.* [9] that propose a multi-body segmentation framework with a direct hill climbing approach that alternates the estimation of region segmentation, camera motion, and depth. This results in a computationally heavy batch algorithm that requires a quite large number of frames to become feasible. Our paper shows that by recovering depth with stereo laparoscopy the problem is considerably simplified, and the region segmentation and camera motion estimation can be performed online as new data arrives.

This paper presents a solution to effectively segment non-rigid or piecewise rigid structures from the surgical site by using multi-model fitting [10]. To solve for the camera relative pose, we use a temporal clustering scheme to better distinguish which scene part should be used to anchor the camera motion estimation. When compared with the state-of-the-art in previously proposed solutions, our method does not require the entire scene to be rigid at an early inspection phase [11], being robust to parts that undergo non-rigid deformation while avoiding priors on these deformations [6]. Quantitative validation is performed with synthetic data [1] to illustrate the numerical stability and performance of the proposed method when the camera motion is accurately known. Qualitative validation in a long *in-vivo* video sequence shows that the proposed method is more effective in recovering the camera motion that the RANSAC-based state-of-the-art in stereo visual odometry [12].

2 Methods

Our method can be split in three main steps: (i) computing dense correspondences between two consecutive images; (ii) generating motion hypothesis using clustering of the motion field with a multi-model fitting approach; (iii) temporal consistency based segmentation of rigid structures that enable the recovery of the camera motion. These steps are described in detail in the sections below.

2.1 Disparity Computation and Pixel-to-Pixel Association

The stereo endoscopic images are assumed to be rectified for disparity map computation and the device is calibrated to determine the intrinsic and extrinsic camera parameters. Given a point $\mathbf{x}_l = (x_l, y_l)^{\mathsf{T}}$ on the left image I_l, the goal is to compute the projection of the same point on the right image I_r that is given by $\mathbf{x}_r = (x_l + d, y_l)$. Ideally, the disparity map D is built by computing d for every image pixel. For the disparity map computation we use the method proposed by Geiger *et al.* [2] that builds a prior over the disparity space by forming a triangulation on a set of robustly matched support points, and subsequently propagates structure into neighbouring image points.

For associating the disparity maps between two consecutive time instants $\mathsf{D}_l \leftrightarrow \mathsf{D}'_l$ we use a standard optical flow method [13] in 2D image space $\mathsf{I}_l \leftrightarrow \mathsf{I}'_l$. For computational reasons we do not compute the flow for every image pixel with a valid disparity and instead we sample the image space by using an equally spaced grid. Our criteria for sampling the grid is defined as function of image resolution to obtain ≈ 4000 point associations between frames.

2.2 Motion Hypothesis Clustering and Refinement with PEaRL

After computing the putative matches $\mathbf{x}_l \leftrightarrow \mathbf{x}'_l$, the correspondence in 3D space $\mathbf{X} \leftrightarrow \mathbf{X}'$ are obtained by using the corresponding disparity values. For registration of the 3D point clouds we use the absolute orientation method [14]. Because different motions can be present at the surgical non-rigid site, we apply the energy-based PEaRL algorithm for labelling the data points with the corresponding motion [10, 15]. This procedure involves three steps: (i) generate an initial set of motion hypotheses, (ii) inlier classification by using an assigned a label (rigid motion) to the putative matches, and (iii) motion refinement using the discrete label assignment.

We start by generating camera motion hypothesis $\mathsf{T} = \begin{bmatrix}\mathsf{R} \ \mathbf{t}\end{bmatrix}$ by sampling sets of 3 neighbouring points (minimal case for [14]) without repetition. Up to 500 motion hypothesis with support larger than 1% the number of pixels on the sample grid are used. Given the set of motion hypothesis $\mathcal{T}$, the goal is to expand the models and estimate their support. This is achieved by applying PEaRL [10] to minimize the energy function

$$E(\mathbf{T}) = \underbrace{\sum_{\mathbf{x}} \mathcal{D}(\mathbf{x}, \mathsf{T}_{\mathbf{x}})}_{\text{Data cost}} + \underbrace{\lambda \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{N}} w(\mathbf{x},\mathbf{y}) \delta(\mathsf{T}_{\mathbf{x}}, \mathsf{T}_{\mathbf{y}})}_{\text{Smoothness term}} + \underbrace{\beta |\mathcal{T}_T|}_{\text{Label cost}} \ , \tag{1}$$

where $\mathbf{T} = \{\mathsf{T}_{\mathbf{x}} | \mathbf{x} \in \mathbf{P}\}$ is an assignment of rigid motion models to data points $\mathbf{x} = \{\mathbf{x}_l, \mathbf{x}'_l\}$. The data cost term $\mathcal{D}(\mathbf{x}, \mathsf{T}_{\mathbf{x}})$ is the reprojection error [16] that enables to measure the error in 2D, which is more robust than directly compute the data cost in the 3D point clouds [12]. The second terms is a smoothness term that encourages the assignment of the same label (rigid motion) to spatially close points. For each data point $\mathbf{x}$ only its 10 nearest neighbours $\mathbf{y}$ are

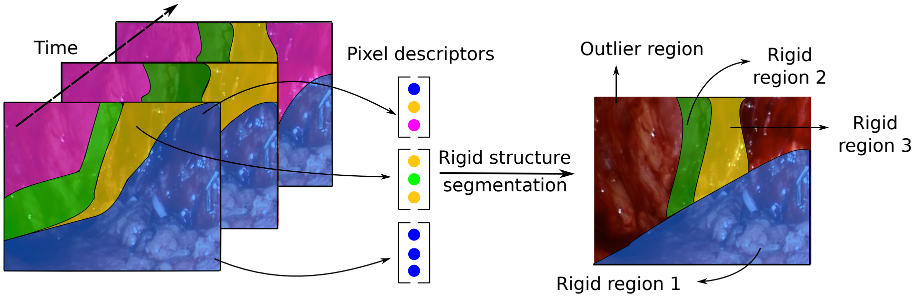


Fig. 1. Rigid segmentation algorithm. At each frame, one label is assigned to a point correspondence (same color represent the same label, and magenta represent the outlier label). While rigid structures tend to be classified with same labels in different views, piecewise rigid or non-rigid parts tend to fragment into different labels or be classified as outliers.

considered to compute the weight $w(\mathbf{x}, \mathbf{p})$. Since we want to enforce spatial consistency in the segmentation we consider that closer points are more likely to be described by the same rigid motion, with the weight being inversely proportional to their euclidean distance. This achieved with the Gaussian function $w(\mathbf{x}, \mathbf{y}) = \exp\left(-||\mathbf{x} - \mathbf{y}||_2/\sigma^2\right)$. $\delta(.)$ represent the Potts model, being 1 when $\mathsf{T}_\mathbf{x} \neq \mathsf{T}_\mathbf{y}$, and 0 otherwise [10, 15]. The label cost penalizes the number of different labels being assigned to the data points to avoid excessive fragmentation. To the possible set of rigid motions $\mathcal{T}$ we add an empty label $\emptyset$, which as a constant data cost of 1.5 pixels for all data point and label cost equal to zero. Occlusions and non-rigid tool-tissue interactions will be intrinsically handled by the outlier. The outlier label also enables to handle erroneous flow estimation and disparity values, avoiding the need to perform the flow (section 2.1) on a temporal window.

After the first label expand, the motion parameters are refined by using the inliers of each label. This is accomplished by minimizing the reprojection error [16] with the Levenberg-Marquardt algorithm [10, 16], with the empty labels being discarded. The new set of labels is then used in a new expansion step with the algorithm iterating between labelling and motion refinement until the optimization does not decrease the energy of Eq. 1 or a certain number of iterations is reached. The constants λ and β were set to $\lambda = 1$ and $\beta = 200$. These values were empirically obtained, and were used across all the experiments.

2.3 Segmenting Multi-View Consistently Labelled Parts

The minimization of the energy function of Eq. 1 guaranties that a label is assigned to each data point $\mathbf{x}$. Since between two consecutive frames the non-rigid or piecewise rigid structures can be subtle and easily confused with the rigid ones, we adopt a window-based system where several frames are used to effectively distinguish between rigid and non-rigid scene parts.

Given a temporal window (see Fig. 1), we build a label-based descriptor for each pixel by concatenating the labels assigned in the frame-to-frame PEaRL optimization. Pixel descriptors with the outlier label assigned in one or more frames are discarded from further processing. The temporal segmentation is carried by clustering pixels with the exact same descriptor. In case of existing more than one cluster, the one with largest spatial support is selected as dominant rigid region and it is used to anchor the relative camera motion. Intuitively, we explore the fact that rigid structures tend to be classified with same labels in different views, while the piecewise rigid or non-rigid parts tend to fragment into different labels or be classified as outliers by the PEaRL algorithm.

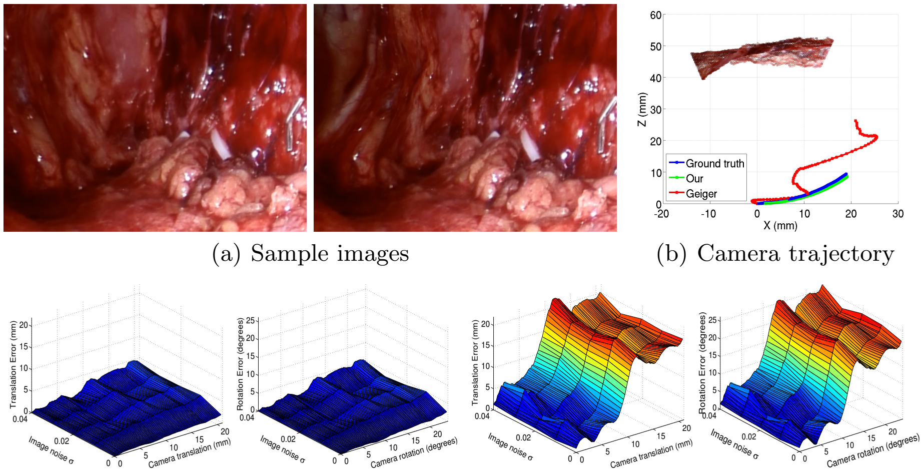
Finally, bundle adjustment [16] is used to refine both the camera motion and the scene structure by using only the dominant rigid part of the scene. This step is necessary because non-rigid regions can contribute on a frame-to-frame basis (locally rigid) to the optimization with PEaRL. Ideally, the temporal segmentation could be computed in automatic manner for adapting to the magnitude of the deformation present in the scene, but this is means deterministic running times would be difficult to guarantee. In practice, we found that the 4/5 frames are sufficient to deal with large deformations and more subtle deformations.

3 Experiments and Results

For validation of the proposed method we conduct experiments with synthetic and *in-vivo* data. The proposed method was fully implemented in MATLAB, with exception of PEaRL which is implemented in C++ code [10]. The single core implementation of the algorithm runs at 0.5 fps in 960 × 540 images on an Intel i7-3630QM CPU @ 2.40GHz processor. Our method is compared with the RANSAC-based approach of [12] which is among the state-of-the art in visual odometry. This method is implemented in C++ and it runs at 2.5 fps after tuning the method parameters to obtain the best possible camera motion estimations.

3.1 Experiments in synthetic data

Camera and scene motion ground truth is difficult to obtain for *in-vivo* MIS video and therefore, the proposed method is validated in a synthetic environment for which the camera motion is precisely known. While simulation sequences cannot render the full complexity of the surgical environment they allow to test the accuracy of the proposed method against different levels of white image noise to illustrate the numerical properties of the method. This sequence comprises 90 frames with the largest part of the scene ($> 60\%$) presenting strong deformation. Figure 2(b) shows the camera motion estimation in the noise free case. It can be seen that our trajectory closely follows the ground truth one, enabling accurate camera estimation in case of such large deformation, while Geiger's method [12] tends to follow the non-rigid deformation motion. Figures 2(c) and 2(d) shows the translation and rotation errors for increasing levels of image noise, showing that our method is numerically stable under moderate amounts of image noise.



(a) Sample images

(b) Camera trajectory

(c) Our method

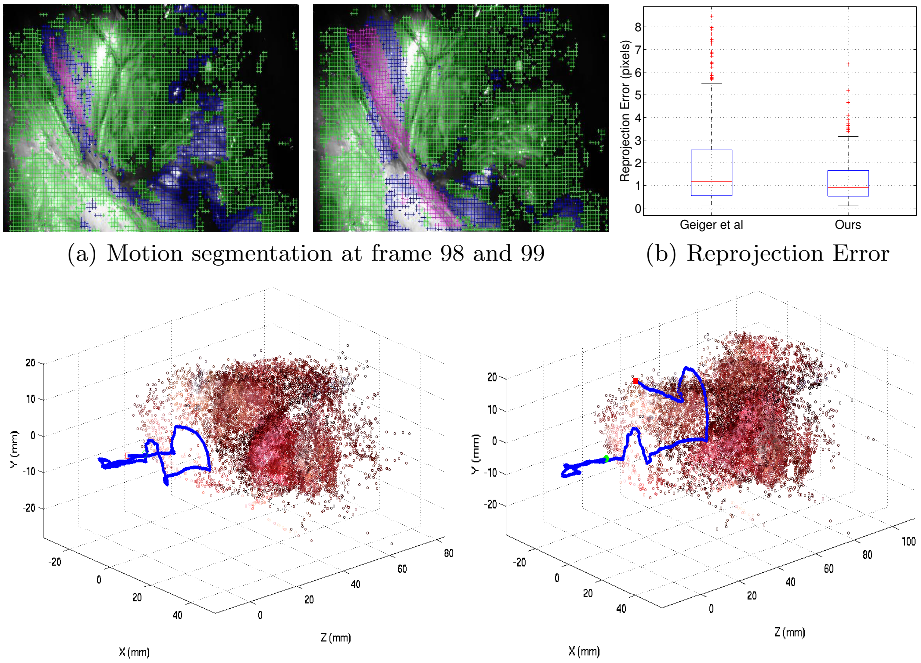
(d) Geiger's method [12]

Fig. 2. Simulation results under increasing level of image noise. (a) show the simulation images with large deformation between them. (b) show the camera trajectory estimation for the zero noise case. Green curve represent the ground truth, blue is ours, and red is obtained with Geiger's method. (c,d) show the performance of both methods under increasing amount of additive white noise. For each method, the left graphics show the translation error as a function of the camera translation motion and level of noise. The same is done for the rotation on the right. It can be seen that our method is numerically stable under moderate levels of image noise.

3.2 Experiments in *in-vivo* data

The data used in this experiment was recorded with *da Vinci Si* surgical robot during a robotically assisted prostatectomy surgery. Our and Geiger's methods [12] were used to recover the camera motion and also the dense 3D scene reconstruction. This sequence of 500 frames is particularly challenging due to the presence of non-rigid motion, strong specularities, bleeding and physiological motion due to large vascular structures in the view. At the end the sequence the camera approximately returns to the starting point performing a loop-closure which can be used for qualitative assessment.

Figure 3 shows the results for camera motion estimation using our and Geiger's methods. Since our solution effectively segments the non-rigid parts of the scene, the camera motion is reliably recovered. Geiger's method employs a conventional frame-to-frame RANSAC-based approach that is less suitable for the challenges in MIS images with the trajectory clearly drifting in the presence of non-rigid motion. To provide a quantitative measure of the quality of the motion estimation, we compute the reprojection error of the reconstruted 3D points, where it can be seen that our method enables more accurate reconstruction and camera motion estimation. It can also be seen that our method is considerably more closer to perform the loop closing, with an error in position



(a) Motion segmentation at frame 98 and 99 (b) Reprojection Error

(c) Our method (d) Geiger's method [12]

Fig. 3. Results in the *in-vivo* sequence. (a) shows two instants with the overlay segmentation. Magenta represent the outlier label that increases with larger deformation. (b) shows the reprojection error obtained with each pixel in frame-by-frame basis. (c,d) show the results of our method and the method of [12] for the camera motion recovery. Our method is capable of performing reliable long-term camera motion estimation, while [12] tends to deteriorate the estimations due to the presence the non-rigid parts.

of 0.6 mm and an orientation error of 5.2 degrees, while the Geiger's method has an error in position of 28 mm and an orientation error of 24.34 degrees.

4 Discussion and Conclusions

We have presented a method for rigid structure segmentation and camera motion estimation during stereoscopic MIS. The proposed method relies on PEaRL [10] for segmenting the scene rigid structures to anchor the camera motion estimation. Temporal consistency is enforced by clustering the segmented scene structures according to the labelling assigned in the PEaRL step. Quantitative and qualitative validation in simulation and *in-vivo* data show that our solution enables to keep accurate camera motion estimation in the presence of significant non-rigid deformation, outperforming a RANSAC-based state-of-the-art method in stereo visual odometry [12]. Future work includes the implementation of our solution for real-time stereo visual odometry using parallelization with GPGPU, and investigation of more suitable solutions for performing the correspondences directly in the 3D space by exploring stereoscopic flow [1].

References

1. Stoyanov, D.: Stereoscopic Scene Flow for Robotic Assisted Minimally Invasive Surgery. In Ayache, N., Delingette, H., Golland, P., Mori, K., eds.: MICCAI. Volume 7510 of LNCS. Springer Berlin Heidelberg (2012) 479–486
2. Geiger, A., Roser, M., Urtasun, R.: Efficient Large-Scale Stereo Matching. In Kimmel, R., Klette, R., Sugimoto, A., eds.: ACCV. Volume 6492 of LNCS. Springer Berlin Heidelberg (2011) 25–38
3. Maier-Hein, L., *et al*: Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. Medical Image Analysis **17** (2013) 974 – 996
4. Roehl, S., , *et al*: Dense gpu-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. Medical Physics **39** (2012) 1632–1645
5. Burschka, D., Li, M., Ishii, M., Taylor, R., Hager, G.: Scale-invariant registration of monocular endoscopic images to ct-scans for sinus surgery. Medical Image Analysis **9** (2005) 413–426
6. Mountney, P., Yang, G.Z.: Motion compensated slam for image guided surgery. In Jiang, T., Navab, N., Pluim, J., Viergever, M., eds.: MICCAI. Volume 6362 of LNCS. Springer Berlin Heidelberg (2010) 496–504
7. Garg, R., Roussos, A., Agapito, L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In: IEEE Conference on Computer Vision and Pattern Recognition. (2013) 1272–1279
8. Luo, X., Mori, K.: Robust Endoscope Motion Estimation Via an Animated Particle Filter for Electromagnetically Navigated Endoscopy. IEEE Transactions on Biomedical Engineering **61** (2014) 85–95
9. Roussos, A., Russell, C., Garg, R., Agapito, L.: Dense multibody motion estimation and reconstruction from a handheld camera. In: IEEE International Mixed and Augmented Reality. (2012) 31–40
10. Isack, H., Boykov, Y.: Energy-based geometric multi-model fitting. International Journal of Computer Vision **97** (2012) 123–147
11. Giannarou, S., Zhang, Z., Yang, G.Z.: Deformable structure from motion by fusing visual and inertial measurement data. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. (2012) 4816–4821
12. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3d reconstruction in real-time. In: IEEE Intelligent Vehicles Symposium. (2011) 963–968
13. Farnebck, G.: Two-Frame Motion Estimation Based on Polynomial Expansion. In Bigun, J., Gustavsson, T., eds.: SCIA. Volume 2749 of LNCS. Springer Berlin Heidelberg (2003) 363–370
14. Horn, B.K.P.: Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America A **4** (1987) 629–642
15. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 1222–1239
16. Triggs, B., Mclauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle adjustment a modern synthesis. In: Vision Algorithms: Theory and Practice, LNCS, Springer Verlag (2000) 298–375