Stereoscopic Scene Flow for Robotic Assisted Minimally Invasive Surgery

Danail Stoyanov

Centre for Medical Image Computing
University College London, WC1E 8BT, UK
{danail.stoyanov}@ucl.ac.uk
http://cmic.cs.ucl.ac.uk

Abstract. Information about the 3D shape and motion of tissue surfaces at the surgical site during minimally invasive surgery is important for providing metric measurements that enable the deployment of image-guidance and enhanced robotic control. This article presents a scene flow algorithm that recovers the deformation and 3D structure of the surgical field-of-view from stereoscopic images by propagating information starting from a sparse set of candidate seed matches. By imposing spatial and temporal constraints the proposed algorithm is able to reconstruct dense 3D scene flow accurately and efficiently. Validation is performed using simulation data to evaluate the method against varying levels of image noise and results are also presented for benchmark phantom model data. The practical value of proposed method is shown by qualitative results for *in vivo* videos from robotic assisted procedures.

1 Introduction

Real-time information about the motion and 3D structure of the surgical site during Minimally Invasive Surgery (MIS) is important for enabling computer assisted interventions and robotic surgical systems with advanced capabilities for navigation and active control [1-4]. With robotic surgical systems, such as da Vinci® by Intuitive Surgical Inc., a stereoscopic laparoscope is used to provide the surgeon with depth perception of the operating field-of-view. The same stereo imaging device can also be used to compute real-time metric measurements from the surgical site using optics and without introducing additional hardware into the patient or the operating theatre [1,2,4]. However, vision-based shape reconstruction and motion tracking are challenging problems due to dynamics at the surgical site and large scale tissue deformation, occlusions from the surgical instruments and the complex scene illumination.

The feasibility of optical 3D reconstruction of the operating field using stereoscopic laparoscopes and computational stereo has previously been reported [5-7]. Preliminary validation studies on phantom models with ground truth data have shown promising results [6] but more comprehensive experimental analysis in complex scenes with realistic tissue reflectance need to be performed. Real-time performance reaching video frame rates for standard resolution images has also been reported [7]. Other optical systems that use active illumination such as structured light [8] and time-of-

2012.

flight [9] have been demonstrated as promising especially when tissue surfaces are homogeneous [1]. These approaches compute a 3D reconstruction of the surgical site but do not retrieve any information about the temporal motion of tissues or instruments. Methods for combined temporal tracking and 3D reconstruction have been reported either using sparse salient features [10,11] or by using parametric surface models of the soft-tissue [12]. While such methods can operate in real-time and naturally enforce surface constraints on the tissue, it is not clear how they can accommodate large occlusions or surface discontinuities between instrument and tissue boundaries. A different approach to dense motion estimation has been investigated with monocular images by using optical flow estimation particularly for deriving the camera pose in diagnostic endoscopy [13-15]. With stereo laparoscopes the optical flow approach can be extended to 3D by computing the flow in both the left and right images and simultaneously estimating the stereo disparity [16, 17].

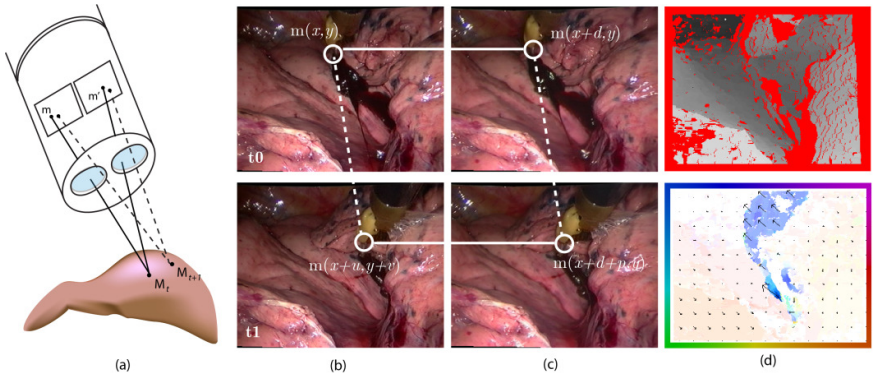

Fig. 1. (a) Schematic illustration of the stereo-laparoscope imaging a moving point on the tissue surface at two time instants; (b-c) rectified stereoscopic images obtained at two time points illustrating the constraints on scene flow motion in the images; (d) example depth map (lighter shade is closer to the camera and blue pixels are occluded) computed from the stereo pair in the top row and the optical flow image computed using scene flow between two time instants shown (color intensity represents magnitude of motion with white being no motion and deep color meaning more, the hue represents the direction as shown around the borders of the image.

In this study, we reports an algorithm for recovering the 3D scene flow at the surgical site by propagating information around a sparse set of corresponding points to both estimate the stereo disparity at each time frame and the temporal motion between consecutive frames. The advantage of this local technique is that it is easy to incorporate constraints on instrument motion and view invariant masking of highlights which can influence global optimization approaches. To the authors' knowledge this is the first work to report 3D scene flow in MIS where both the motion and the structure of the operating field are recovered in 3D. Validation using synthetic data and phantom models illustrates the performance of the method and qualitative results on *in vivo* videos from robotic assisted surgical procedures indicate that the method can potentially be used in clinical practice. An executable of the simulation environment used

to generate synthetic validation data and the source code for algorithm reported in this study are available online[1].

2 Methods

This article reports a novel method for determining the 3D structure of the surgical site and its temporal motion. The technique involves determining the disparity at each time frame, estimating the 2D optical flow in the left and right views and subsequently determining a consistent 3D flow field and detecting occluded regions.

2.1 Disparity estimation

The stereo laparoscope is assumed to be calibrated such that the intrinsic and extrinsic camera parameters are known and the toolbox used to perform the calibration is available online[1]. For each incoming stereoscopic image pair at time t the images are rectified to remove lens distortions and to align the epipolar geometry by using the known calibration parameters of the cameras [18]. From the rectified images the disparity $d(x,y,t)$ at an image pixel in the left image $\mathbf{m}_t^l = [x,y]^T$ provides the correspondence to the projection of the same world point in the right image as $\mathbf{m}_t^r = [x + d(x,y,t), y]^T$. The disparity map is estimated at each time frame by using the implementation of the algorithm in [6] which is also available online. This is based on a growing scheme [19] from an initial set of seed points that are matched across the stereoscopic view using a sparse matching algorithm [10]. The search space for growing is restricted to 1D by rectification and a symmetry constraint is added to ensure left-right disparity map consistency. We estimate the disparity map at every frame in order to decouple the flow and disparity computations and optimize each problem individually as has been reported to be effective for scene flow [17].

Any feature point detection and feature matching strategy can be used to generate seed points for the disparity growing scheme. We use simple corner features based on the image gradients as they can be computed efficiently and have previously been shown to work well for short-term tracking in MIS images with a stereoscopic tracking method [10]. More complex strategies and feature detectors or descriptors can be adapted to work within the proposed framework at the cost of additional computational load.

2.2 Scene Flow Estimation

The idea of scene flow is illustrated in Fig 1 where $\mathbf{m}_t^l$ and $\mathbf{m}_t^r$ are the pixel projection coordinates in the left and right stereo images of a point on the tissue surface $\mathbf{M}_t = [X,Y,Z]^T$ at time t . At time $t+1$ the point in the left image $\mathbf{m}_{t+1}^l$ corresponding to $\mathbf{m}_t^l$ can be written as $\mathbf{m}_{t+1}^l = [x + u(x,y,t), y + v(x,y,t)]^T$ and similarly for the right image the point corresponding to $\mathbf{m}_t^r$ can be written as $\mathbf{m}_{t+1}^r = [x + d(x,y,t) + p(x,y,t), y + v(x,y,t)]^T$. In 2D image space the optical flow field for the left image is defined by $[u(x,y,t), v(x,y,t)]^T$ but because we have stereo-

[1] http://www.cs.ucl.ac.uk/staff/dan.stoyanov/software.html

scopic information we can derive the full scene flow for the 3D motion defined by $[u(x,y,t), v(x,y,t), p(x,y,t)]^T$ where the term $p(x,y,t)$ represents the change in disparity between t and $t+1$. By computing the parameters $[u,v,p]^T$ (omitting image and time notation for clarity) we can calculate the full 3D scene flow.

For an incoming stereo image pair, given the disparity map generated at the previous time frame with the method in Section 2.1 we can make the several measurements to compute the flow information by measuring the similarity between image regions, we define:

$$\varepsilon_{ll} = \Theta\left(\mathbf{m}_t^l, \mathbf{m}_{t+1}^l\right) \quad \varepsilon_{rr} = \Theta\left(\mathbf{m}_t^r, \mathbf{m}_{t+1}^r\right) \quad \varepsilon_{lr} = \Theta\left(\mathbf{m}_{t+1}^l, \mathbf{m}_{t+1}^r\right) \qquad \textbf{(1)}$$

Where the similarities of image regions denoted by ε are determined by the function Θ which is the zero mean normalized cross correlation measured between rectangular image windows centered at each point of interest. Using these measures it is possible to formulate the scene flow problem within a variation framework [13], however, this imposes smoothness priors that can be problematic in occluded areas or in regions with specular reflection. We therefore use a growing scheme similar to the one used for stereo matching in Section 2.1 and originally developed in [19] and recently adapted for scene flow in urban environments [16].

Starting from the set of candidate seed matches computed in Section 2.1 for both disparity and temporal motion we propagate information around each match using the best-first principle. The seeds are stored in a priority queue determined by their similarity scores from (1) and therefore obtaining the best seed to propagate at each step is performed by popping the queue. We perform the propagation independently in the left and right channels, which may seem redundant, but we exploit the redundancy to perform consistency and symmetry checking thus detecting occlusions in the flow as well as in the disparity. Furthermore, because the propagation is constrained by the epipolar geometry and by a disparity smoothness threshold, which we limit to one, there is an overlap of correlation computations which we can exploit for efficiency. Finally, we run the algorithm hierarchically starting with small correlation windows and then repeating with larger ones but using the earlier result as an initialization seed priority queue. The rejection scheme handles error propagation naturally in this case and the larger windows are able to fill in homogeneous regions more reliably.

3 Experiments and results

The proposed method was implemented using C++ and, without specific optimization or parallelization, it is able to operate at approximately 1Hz for 360 x 288 images on a single core of an Intel i7 M620 2.76GHz mobile processor. For our simulation validation studies we used a custom simulation environment where textures are used with a surface model that can be augmented to simulate tissue deformations induced by the cardiac cycle and respiration. The environment is available online[2] and has a number of parameters that can be used to customize the virtual cameras, the amount of additive Gaussian noise and the type of deformation induced on the surface. We also re-

[2] http://www.cs.ucl.ac.uk/staff/dan.stoyanov/software.html

port results for the heart phantom datasets reported in [6] and made available by the Hamlyn Centre, Imperial College London[3]. Finally we show qualitative results on the *in vivo* data made available in [1,10].

3.1 Experiments with synthetic data

Ground truth information for 3D scene flow is not available in surgery and even for phantom experiments linking the temporal motion of dense surface points is not currently possible. Therefore we evaluate the stability and performance of the proposed method on synthetic data with varying levels of additive Gaussian noise with zero mean and increasing standard deviation. While simulation environments cannot render a fully photorealistic representation of the surgical site they allow testing the robustness of an approach against known ground truth information.

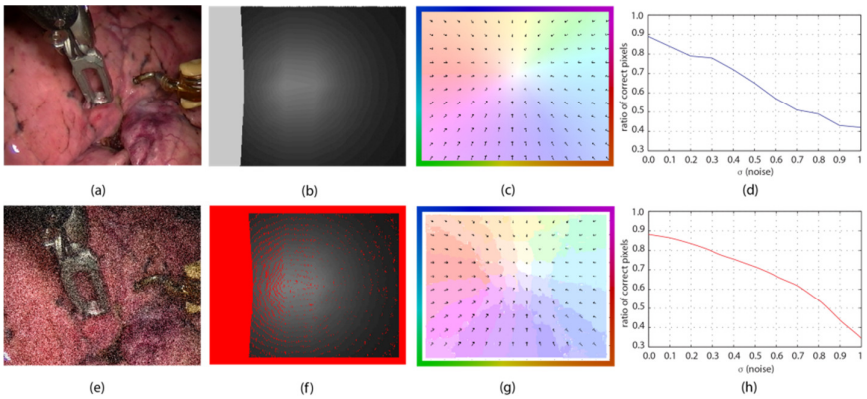

Fig. 2. (a,e) Images generated from the simulation environment with and without additive noise; (b, f) ground truth disparity map from the simulation and below the disparity generated with the proposed technique; (c, g) ground truth optical flow map in the left image (sub-pixel) and below the computed flow map using the proposed method; (d, h) plot of the error for disparity computation (blue) and optical flow (red) against varying levels of additive noise.

The results shown in Fig 2(g, h) indicate that the proposed method performs well on the synthetic data. We show ground truth disparity and flow images in Fig 2(b, c) and the corresponding example reconstructed disparity map in Fig 2(f, g) when additive image noise has been introduced to the image as shown in Fig 2(e). It is clear that there is good agreement between the ground truth and our results, however, our method operates only on integer values and therefore cannot match the sub-pixel quality of the ground truth. This results is banding of the results visible in the images but can be removed with a final subpixel refinement step. The plots in Fig 2(d) show the performance of our method against varying levels of noise where the deviation of additive noise is normalized in the 0-1 range.

[3] http://hamlyn.doc.ic.ac.uk/vision/

3.2 Experiments on phantom model data

To evaluate the method proposed in this study against phantom model data we used the heart model data reported in [5]. The two datasets are of a beating heart phantom model with ground truth obtained using dynamic CT scanning to measure the geometry of the model. The data does not have temporal connectivity available and therefore evaluating the 3D scene flow we compute is not possible for this data. Hence we only compare the disparity results obtained with our technique to the ground truth disparity at each frame in the video sequences averaged over one cardiac cycle.

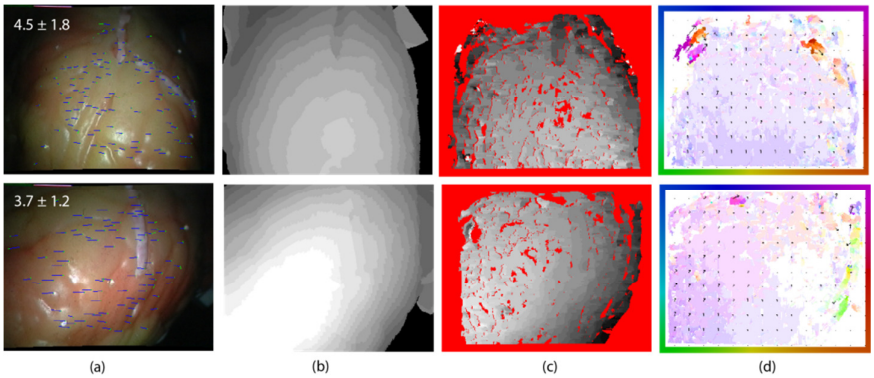

Fig. 3. (a) Two example images with seed feature tracks matched in stereo shown in blue and temporal tracks shown in green; (b) ground truth disparity images obtained from CT data; (c) corresponding disparity images obtained with our method; (d) the flow images corresponding to inter-frame motion.

Fig 3 shows example images from the heart phantom with sparse feature tracks used to initialize our method overlaid on top of the video. The disparity maps resulting at a time frame generated by the proposed technique are shown in Fig 3(c) and visibly correspond well to the ground truth data. We ran our technique over both video sequences for a full cardiac cycle and the resulting disparity error and deviation are overlaid in Fig 3(a). The disparity error for each dataset was measured as 4.5±1.8 pixels and 3.7±1.2 pixels. The flow information shown has no ground truth but intuitively we observe larger motion close to the camera as the heart model simulates a cardiac cycle. While our errors are higher than reported in [6] it is important to note that we are computing disparity over the entire cycle of heart data and not on a single frame. This has a disadvantage because the video and dynamic CT data are not perfectly aligned in time and a conversion formula is used (please see the data's website).

3.3 Experiments with *in vivo* data

We illustrate the practical value of the method proposed in this article by applying to several videos taken in vivo during robotic assisted surgery. The results shown in Fig 4 clearly capture the visual appearance of the 3D structures within the scene and per-

form well in terms of not mismatching occluded regions even with the presence of large instruments in the foreground.

It is more difficult to visualize the reconstructed 3D motion but qualitatively we can see that it corresponds to instrument motion where present and to different tissue surface planes in the scene. The motion data is best visualized using the video submitted as supplementary material for this submission. Naturally some errors are apparent and in the 3D reconstruction these are usually due to sharp discontinuities meanwhile in the motion fields they typically reflect sudden changes in motion direction

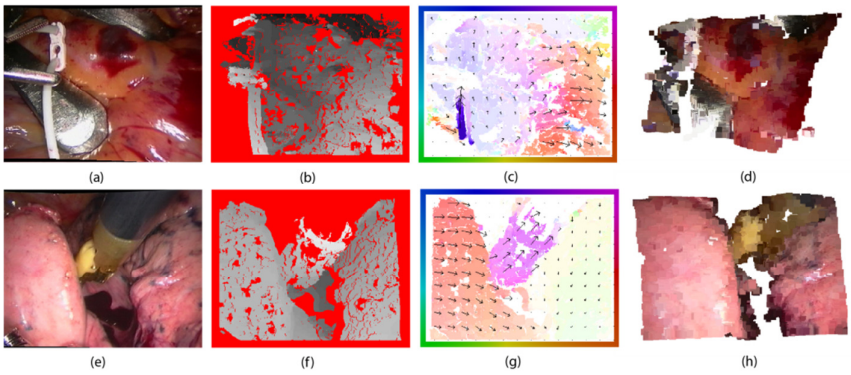

Fig. 4. (a, e) Example stereoscopic images from endoscopic beating heart surgery and a robotic procedure on the lung; (b, f) disparity images showing the 3D shape and (c, g) flow dynamics of the operating field; (d-h) renderings of the recovered 3D geometry of the surgical site without incorporated occluded regions.

4 Discussion

In this article, we have presented stereoscopic framework for recovering the 3D structure and motion of the operating field during robotic assisted MIS. The method is robust as it uses a growing scheme that rejects outliers and ensures uniqueness and symmetry in the resulting disparity and flow estimates. We have shown that the method performs well against additive image noise on synthetic data and on benchmark phantom model data with known ground truth. Qualitative experiments on *in vivo* datasets from robotic assisted surgery also suggest that the method has practical value. We believe the method is capable of real-time performance with suitable code optimization and a hardware implementation utilizing parallelization. Furthermore the approach can be improved to provide subpixel results with a final refinement step. Our future work will focus on improving the computational performance of the technique and also on investigating more optimal propagation strategies with learned priors, occlusion boundaries and instrument detection.

Acknowledgements. We would like to acknowledge the data provided by Prof Guang-Zhong Yang and the Hamlyn Center, Imperial College London. This work was supported by a Royal Academy of Engineering/EPSRC Research Fellowship.

References

[1] Mountney, P. *et al.*: Motion Compensated SLAM for Image Guided Surgery. In: Jiang, T., Navab, N., Pluim, J. P. W., Viergever, M. A. (eds.) MICCAI 2010, LNCS, vol. 6362, pp. 496-504 (2010)

[2] Mirota, D. J., *et al.*: Vision-Based Navigation in Image-Guided Interventions. Ann. Rev. Biomed. Eng. 13, 297-319 (2011)

[3] Hager, G. *et al.*: Surgical and interventional robotics: part III [Tutorial]. IEEE Robot. Autom. Mag. 15, 84-93 (2008)

[4] Stoyanov, D.: Surgical Vision. Ann. Biomed. Eng. 40, 332-34 (2012)

[5] Devernay, F., *et al.*: Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery. In: *MIAR*, 2001

[6] Stoyanov, D*., et al.*: Real-Time Stereo Reconstruction in Robotically Assisted Minimally Invasive Surgery. In: Jiang, T., Navab, N., Pluim, J. P. W., Viergever, M. A. (eds.) MICCAI 2010, LNCS, vol. 6362, pp. 275-282 (2010)

[7] Röhl, S., *et al.*: Dense GPU-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. Med. Phys. 39, 1632-45 (2012)

[8] Clancy, N. T., *et al.*: Spectrally encoded fiber-based structured lighting probe for intraoperative 3D imaging. Biomed. Opt. Express. 11, 3119-3128 (2011)

[9] Penne, J., *et al.*: Time-of-Flight 3-D Endoscopy. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, LNCS, vol. 5761, pp. 467-474 (2009)

[10] Stoyanov, D. *et al.*: Soft-tissue Motion Tracking and Structure Estimation for Robotic Assisted MIS Procedures. in In: Duncan, J. S., Gerig, G. (eds.) MICCAI 2005, LNCS, vol. 3749, pp. 139-146 (2005)

[11] Paul, P., *et al.*: A Surface Registration Method for Quantification of Intraoperative Brain Deformations in Image-Guided Neurosurgery. IEEE Trans. Inf. Tech. Biomed. 13, 976-983 (2009)

[12] Richa, R., *et al.*: Towards robust 3D visual tracking for motion compensation in beating heart surgery. Med Image Anal. 15, 302-315 (2011)

[13] Deguchi, D. *et al.*: New Image Similarity Measure for Bronchoscope Tracking Based on Image Registration. In: Ellis, R. E., Peters, T. M. (eds.) MICCAI 2003, LNCS, vol. 2878, pp. 399-406 (2003)

[14] Jianfei, L., *et al.*: A stable optic-flow based method for tracking colonoscopy images. In: CVPRW. 1-8 (2008)

[15] Mori, K., *et al.*: Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images. Med Imag Anal. 6, 321-336 (2002)

[16] Cech, J., *et al.*: Scene flow estimation by growing correspondence seeds. In: CVPR. 3129-3136 (2011)

[17] Wedel, A., *et al.*: Stereoscopic Scene Flow Computation for 3D Motion Understanding. Int. J. Comp. Vis. 95, 29-51 (2011)

[18] Hartley, R. and Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge Press (2000)

[19] Lhuillier, M., *et al.*: Robust dense matching using local and global geometric constraints. In: ICPR. 1968-1972 (2000)