Stereo and Shape-from-Shading Cue Fusion for Dense 3D Reconstruction in Endoscopic Surgery

Marco Visentini-Scarzanella, *Member, IEEE,* and Danail Stoyanov, *Member, IEEE*

***Abstract*—Dense 3D reconstruction of the surgical site is important for providing image-guidance, augmented reality and, in robotic surgery, active constraints. The challenge with endoscopic images is that soft-tissue surfaces do not always have salient characteristics and computational techniques fail to achieve unique correspondence. In this paper, we propose a novel method for handling homogenous regions by fusing visual cues using a combination of Shape-from-Shading (SFS) and stereo. A sparse reconstruction of the underlying structure is performed with a feature-based algorithm and used to initialise and guide two independent SFS modules, which infer monocular dense relative depth. The two reconstructions are then registered with nearest neighbour 3D matching, which is directly translated into a dense 2D disparity estimate. Our only assumption is consistency of the two reconstructions and this is sufficient for the overall scheme to be effective. We validate the approach quantitatively with benchmark phantom data and comparison against the state-of-the-art endoscopic reconstruction algorithms.**

I. Introduction

RECONSTRUCTING dense 3D information from intra-operative endoscopic videos is a fundamental building block for computer-assisted endoscopic interventions including image-guided navigation and dynamic active constraints [1], [2]. Active techniques such as structured light, time-of-flight and photometric methods can recover depth information without complex computational algorithms but they require significant adaptation to instrumentation which is a barrier to clinical translation. Vision-based techniques are currently an attractive approach for recovering the *in vivo* tissue morphology since they rely only on image information that is inherently used by the surgeon during procedures. The challenge of vision approaches is that surgical scenes are non-rigid, dynamic, have complex reflectance and surgical instruments create large discontinuities and occlusions.

Different visual cues can be used for reconstructing 3D geometry in endoscopic images [1], [3]. Stereoscopic approaches have received particular attention recently [4], [5] because stereo-laparoscopes are part of robotic surgery systems and are also now available for conventional laparoscopic instruments. One advantage of computational stereo is that tissue deformation due to breathing, peristalsis or the cardiac cycle can be dealt with on a per frame basis. This overcomes some of the difficulties in non-rigid Structure-from-Motion techniques that solve a complex and computationally intensive reconstruction problem [6]. While local propagation based computational stereo techniques have been shown to yield promising results for recovering 3D and motion [4], [3] such methods rely on salient tissue textures and produce results dependent on aggregation or correlation window sizes. Shape-from-Shading (SFS) on the other hand does not require textured surfaces but relies on homogeneous intensity variation to infer dense depth and the surface normals [7], [8], [9], [6], [10]. Recent applications of SFS to endoscopic data have attempted to use specularities to register the reconstructed surface in metric space [7], recovering the surface albedo through pre-operative calibration procedures [9], and methods to model the light source and estimate non-Lambertian BRDF parameters [10], [6] have been presented. Nevertheless, the numerical sensitivity of SFS makes it unsuitable for systems relying exclusively on it for reconstruction [10]. Attempts to combine the complementary attributes of stereo and SFS in endoscopy have been reported [11], [12] but were limited to local SFS assuming orthographic projection and suffered from cumulative errors because of the gradient propagation strategy. However, fusing visual cues is critical for building robust algorithms that can perform reliably like the human visual system even in difficult surgical images.

In this paper, we propose a novel method for combining stereo and SFS reconstruction to robustly recover dense 3D surface information from endoscopic videos. We exploit stereo vision by accurately reconstructing salient areas with strong texture and use this information to provide an initial estimate of the tissue albedo. Then each independent monocular channel is recovered via SFS reconstructions and aligned within the same frame of reference, providing estimates for textureless areas not recovered from stereo. Given the assumption that SFS is consistent between the two channels, which is strengthened further the common stereo initialisation of the two reconstructions, the two dense SFS reconstructions are registered together and their voxels matched in 3D. We validate our method quantitatively with phantom data and measure performance against image noise comparing it to state-of-the-art dense stereo reconstruction algorithms.

II. Method

The proposed method is schematically presented in Figure 2. Given a stereo image pair (I_L, I_R) from a calibrated laparoscope (C_L, C_R) images are fed into a sparse reconstruction algorithm **(1)** [4]. Each image is then processed independently by a SFS module together [7] with the sparse depth information **(2)**. The sparse stereo reconstruction initialises

M. Visentini-Scarzanella is with the Communications and Signal Processing Group, Imperial College London, London, UK, e-mail: (see http://www.commsp.ee.ic.ac.uk/ marcovs/contacts/).

D. Stoyanov is with the Centre for Medical Image Computing, University College London, UK, e-mail: (see http://www0.cs.ucl.ac.uk/staff/Dan.Stoyanov/).

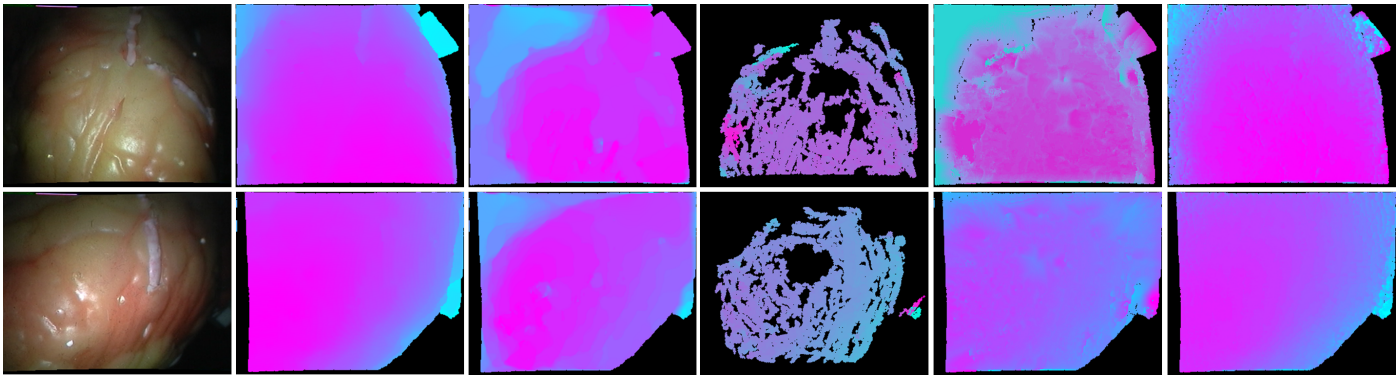

Fig. 1: From left to right: frame from the 'f5' and 'f7' datasets, ground truth disparity, disparities from the TV-L1 [13] algorithm, initial stereo reconstruction [4], SFS reconstruction initialised with stereo data, SFS reconstruction initialised with 5% of ground truth data.

depth and local albedo information for propagation by the SFS module to the remaining pixels. The two SFS reconstructions are then projected in 3D space and registered together for matching **(3)** with the ICP algorithm. As long as the SFS reconstructions are consistent between the two views, they should align perfectly after registration. This allows voxel matching from the two reconstructions via nearest neighbour search and immediately translates to matching between pixels through calibrated projection. A final step can be performed to optimise pixel matches with sub-pixel accuracy and these can be triangulated for the final 3D reconstruction.

III. Results

We evaluate our approach quantitatively on phantom data from the Hamlyn dataset (http://hamlyn.doc.ic.ac.uk/vision/). This consists of two stereo videos with ground truth from aligned dynamic CT data. We test the performance of the algorithm in two different scenarios: by initialising the SFS with different proportions of ground truth data and by initialising the SFS with the sparse stereo reconstruction instead. We further compare our approach with the dense stereo TV-L1 algorithm [13].

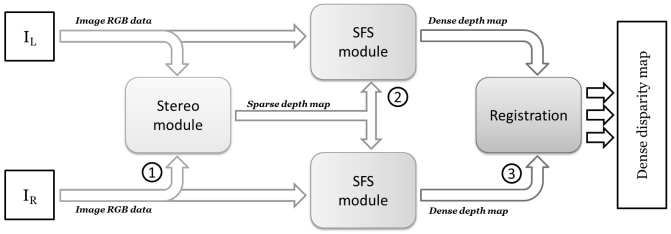

Fig. 2: Workflow for the proposed algorithm. (**1**) An input stereo image pair (I_L, I_R) is used to obtain an initial sparse stereo reconstruction. (**2**) The obtained depth map is then used with each image for a guided Shape-from-Shading dense depth estimate. (**3**) The two depth maps are then registered together for dense disparity map estimation.

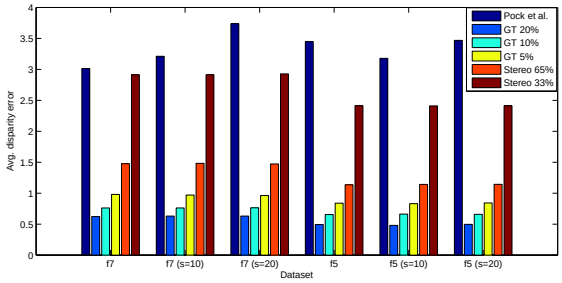

Fig. 3: Average pixel disparity error for the proposed method with different proportions of ground truth (GT) or stereo seeds and state-of-the-art dense stereo [13].

As shown in Figure 3, the ground truth initialisation allows the SFS algorithm to perform well even with very low proportions of initialised elements. The average disparity error was found to be below 2 pixels whenever more than 1% of the total image pixel count is supplied as seeds. The disparity error was measured exclusively on the reconstructed pixels. When the algorithm is initialised by very sparse stereo data, the performance degrades, since even with relatively accurate 3D information the surface normals estimated from stereo data for the albedo calculations are significantly different from the ground truth data. This can be handled by outlier rejection schemes and our current work does not incorporate this and wrong initial stereo matches can be propagated to the dense depth map. Nevertheless, our results show the capabilities of the proposed algorithm even in the presence of such initialization errors compared to the state-of-the-art. Indeed, global variational methods such as [13] suffer from the general lack of texture typical of endoscopic data, and the smoothness constraints imposed by such schemes confers the final disparity map a characteristically blotchy appearance. A visual representation of the dense disparities for the phantom and on *in vivo* data is shown in Figure 1.

IV. Conclusion

Robust reconstruction algorithms that can perform reliably, like the human visual system, in difficult surgical scenes are important for surgical vision and visual cue fusion is a critical element in realising robustness. Here we have proposed a method for 3D tissue surface reconstruction in endoscopic surgery using a novel combination of stereo and SFS. We have reported validation results on benchmark phantom data and compared our method against the state-of-the-art in endoscopic scene 3D reconstruction. In our future work we will focus on *in vivo* experimentation and on building discontinuity models into the framework so that instruments and occlusions do not perturb our reconstruction results.

References

[1] P. Mountney and G.-Z. Yang, "Motion compensated slam for image guided surgery," in *Medical Image Analysis and Computer-Assisted Intervention (MICCAI)*, Beijing, China, 2010, pp. 496–504.

[2] D. J. Mirota, M. Ishii, and G. D. Hager, "Vision-based navigation in image-guided interventions," *Annual Review of Biomedical Engineering*, vol. 13, no. 1, pp. 297–319, 2011.

[3] D. Stoyanov, "Stereoscopic scene flow for robotic assisted minimally invasive surgery," in *Medical Image Analysis and Computer-Assisted Intervention (MICCAI)*, Nice, France, 2012, pp. 479–486.

[4] D. Stoyanov, M. Visentini-Scarzanella, P. Pratt, and G.-Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Beijing, China, 2010, pp. 275–282.

[5] S. Röhl, S. Bodenstedt, S. Suwelack, and *et al.*, "Dense gpu-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration," *Medical Physics*, vol. 39, no. 3, pp. 1632–1645, 2012.

[6] A. Malti and A. Bartoli, "Estimating the cook-torrance brdf parameters in-vivo from laparoscopic images," in *Workshop on Augmented Environment in Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Nice, France, 2012.

[7] M. Visentini-Scarzanella, D. Stoyanov, and G.-Z. Yang, "Metric depth recovery from monocular images using shape-from-shading and specularities," in *IEEE International Conference on Image Processing (ICIP)*, Orlando, USA, 2012, pp. 25 –28.

[8] C. Wu, S. Narasimhan, and B. Jaramaz, "A multi-image shape-from-shading framework fornear-lighting perspective endoscopes," *International Journal of Computer Vision*, vol. 86, pp. 211–228, 2010.

[9] G. Ciuti, M. Visentini-Scarzanella, A. Dore, A. Menciassi, P. Dario, and G.-Z. Yang, "Intra-operative monocular 3d reconstruction for image-guided navigation in active locomotion capsule endoscopy," in *IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 768 –774.

[10] T. Collins and A. Bartoli, "Towards live monocular 3d laparoscopy using shading and specularity information," in *International Conference on Information Processing in Computer-Assisted Interventions (IPCAI'12)*, Pisa, Italy, 2012, pp. 11–21.

[11] M. Visentini-Scarzanella, G. P. Mylonas, D. Stoyanov, and G.-Z. Yang, "*i*-brush: A gaze-contingent virtual paintbrush for dense 3d reconstruction in robotic assisted surgery," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, London, UK, 2009, pp. 353–360.

[12] B. P. L. Lo, M. Visentini-Scarzanella, D. Stoyanov, and G.-Z. Yang, "Belief propagation for depth cue fusion in minimally invasive surgery," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, New York, USA, 2008, pp. 104–112.

[13] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, "Global solutions of variational models with convex regularization," *SIAM Journal on Imaging Sciences*, vol. 3, no. 4, pp. 1122–1145, 2010.