Real-Time Dense Stereo Reconstruction Using Convex Optimisation with a Cost-Volume for Image-Guided Robotic Surgery

Ping-Lin Chang[1], Danail Stoyanov[3], Andrew J. Davison[1], and
Philip "Eddie" Edwards[1,2]

[1] Department of Computing
[2] Department of Surgery and Cancer
Imperial College London, United Kingdom
{p.chang10, a.davison, eddie.edwards}@imperial.ac.uk
[3] Centre for Medical Image Computing and Department of Computer Science
University College London, United Kingdom
danail.stoyanov@ucl.ac.uk

Abstract. Reconstructing the depth of stereo-endoscopic scenes is an important step in providing accurate guidance in robotic-assisted minimally invasive surgery. Stereo reconstruction has been studied for decades but remains a challenge in endoscopic imaging. Current approaches can easily fail to reconstruct an accurate and smooth 3D model due to textureless tissue appearance in the real surgical scene and occlusion by instruments. To tackle these problems, we propose a dense stereo reconstruction algorithm using convex optimisation with a cost-volume to efficiently and effectively reconstruct a smooth model while maintaining depth discontinuity. The proposed approach has been validated by quantitative evaluation using simulation and real phantom data with known ground truth. We also report qualitative results from real surgical images. The algorithm outperforms state of the art methods and can be easily parallelised to run in real-time on recent graphics hardware.

1 Introduction

An important challenge in robotic-assisted laparoscopic surgery is the 3D reconstruction of the observed surgical site. The recovered 3D scene can provide a rich source of information for visualisation and interaction, enabling vision-based camera tracking and registration to a preoperative model for surgical navigation [2, 6, 7]. With the da Vinci surgical system the presence of a stereoscopic laparoscope means that computational stereo is a practical and feasible approach to *in vivo* reconstruction [3, 15]. However, surgical scenes are challenging for 3D reconstruction algorithms because of texture-poor appearance, occlusions, specular reflection and discontinuities due to instruments.

This research is partly supported by ERC Starting Grant 210346, CRUK grant A8087/C24250 and The Royal Academy of Engineering/EPSRC Research Fellowship.

Reconstruction of the stereo-endoscopic view for surgical navigation has been an active area of research for over a decade [6,7]. Much of the prior work has focused on beating heart surgery [3,5,9,15], where the reconstructed heart surface could be used for motion stabilisation or registration to a preoperative model. To achieve smooth and robust stereo reconstruction, methods have been proposed that use a parametric surface description [5] to overcome texture homogeneity. Alternatively region growing starting from sparse features has been reported [15] and thin-plate spline interpolation of robust features [9]. A sophisticated framework which uses a hybrid CPU-GPU algorithm to fuse temporal reconstruction into a global model has been proposed [10]. In all cases the aim is to approach real-time reconstruction, and to this end GPU implementations and parallelisation are necessary.

In this paper, we build on recent advances in computer vision and the use of variational techniques to efficiently and effectively reconstruct stereo-endoscopic scenes using stereo image pairs. This is achieved by constructing a cost-volume with a reliable data term and performing convex optimisation to solve a Huber-L^1 model. The proposed algorithm can also be effectively parallelised on the GPU for real-time performance. Compared with the state of the art, the proposed approach yields more accurate reconstruction in empirical studies. We illustrate this with extensive validation using synthetic and phantom data with known ground truth and qualitative results from *in vivo* robotic surgery sequences.

2 Proposed Approach

The first step of the proposed algorithm is to construct a 3D cost-volume using a pixel-wise data term with respect to the disparities. An efficient convex optimisation for solving a Huber-L^1 model is then performed by decoupling the model into a Huber-L^2 model and the cost-volume, which can be resolved by a primal-dual algorithm and exhaustive search alternately.

2.1 Cost-volume construction

In definition, a cost-volume $C : \mathbf{\Omega}_C \to \mathbf{R}$, where $\mathbf{\Omega}_C \subseteq \mathbf{R^3}$, is a discrete function which maps a 3-vector to a cost value. In rectified stereo matching the cost-volume is also called the disparity space image (DSI) [12] which is defined as

$$C\big(\mathbf{x},\, \mathbf{u}(\mathbf{x})\big) = \rho\big(I_l(\mathbf{x}),\, I_r(\mathbf{x}')\big). \tag{1}$$

The stereo images are assumed to be undistorted and rectified in advance. Functions I_l and $I_r : \mathbf{\Omega}_I \to \mathbf{R^3}$ are the left and right colour image and $\mathbf{\Omega}_I \subseteq \mathbf{R^2}$. As per convention, the Eq. 1 takes the left image as reference and stereo matching is performed in the right image, and thus $\mathbf{x} = (x, y)^\top$ and $\mathbf{x}' = (x + \mathbf{u}(\mathbf{x}), y)^\top$. The function $\mathbf{u} : \mathbf{\Omega}_I \to \mathcal{D}$ maps a pixel location to a set of discrete integer disparities within a range $\mathcal{D} = [d_{min}, d_{max}]$. The cost-volume C is then constructed using all of the disparities for each pixel in the image domain $\mathbf{\Omega}_I$. The size of the cost-volume is therefore $|\mathbf{\Omega}_I| \times |\mathcal{D}|$. Note that the resolution of the disparity $|\mathcal{D}|$, d_{min} and d_{max} are dependent on scenes and camera profiles.

Robust data-fidelity term. In a pure vision-based reconstruction problem, the function ρ in the Eq. 1 can be an arbitrary photometric measure which defines the data-fidelity term. The data fidelity is essential since the later convex optimisation significantly relies on it.

We illustrate the effects of different measures with a simulation stereo pair generated by a textured cone model as shown in Fig. 1a. Raw reconstruction is achieved using a winner-takes-all scheme which extracts the disparity pixel-wise according to the minimum cost. The simplest absolute difference (AD) measure gives a very noisy raw reconstruction as shown in Fig. 1b. To reduce the noise, one may adopt the sum of absolute differences (SAD) or sum of squared differences (SSD) which aggregate costs locally. Alternatively, applying more sophisticated edge-preserving local filtering can yield an even better result [8, 10]. Fig. 1c shows the result after bilateral filtering (BF) is applied to Fig. 1b. However, our empirical studies have shown that if the original measure is error-prone, the later aggregation in the cost-volume space can increase the error, which results in poor data-fidelity. This commonly happens in textureless regions, half-occluded areas and where the illumination changes.

In contrast, zero-mean normalised cross-correlation (ZNCC) implicitly performs the aggregation using a window patch, so correlation is calculated over a pixel neighbourhood. This results in a measure more tolerant to different camera gain or bias and can also provide better fidelity in textureless regions. Fig. 1d shows the raw reconstruction using ZNCC. In this work we therefore use ZNCC as the data term measure to construct the cost-volume.

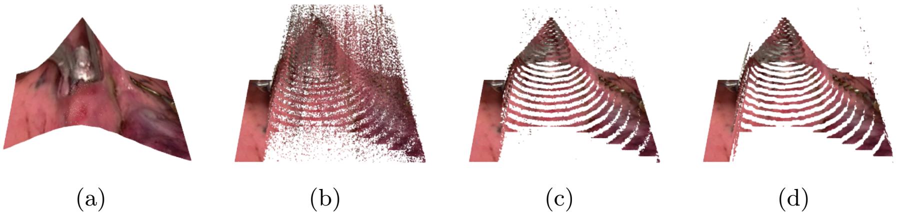

Fig. 1: The raw reconstruction results using a winner-takes-all scheme with the cost-volume. (a) The simulated ground truth textured model. (b) The raw reconstruction using AD. (c) The result after applying BF to (b). (d) The raw reconstruction using ZNCC.

2.2 Huber-L^1 convex optimisation with the cost-volume

Starting with the coarse reconstruction, the unknown disparity function $\mathbf{u}$ is further optimised by solving a Huber-L^1 variational energy functional which takes the cost-volume as data term and an image-driven weighted Huber-norm as a regulariser term. This is defined as

$$E(\mathbf{u}) = \int_{\mathbf{\Omega}_I} \Big\{ w(\mathbf{x}) \|\nabla \mathbf{u}(\mathbf{x})\|_{\varepsilon} + \lambda C(\mathbf{x}, \mathbf{u}(\mathbf{x})) \Big\} \, \mathrm{d}\mathbf{x}, \tag{2}$$

where

$$\| \cdot \|_\varepsilon = \begin{cases} \frac{\|\cdot\|_2^2}{2\varepsilon} & \text{if } \| \cdot \|_2 \leq \varepsilon \\ \| \cdot \|_1 - \frac{\varepsilon}{2} & \text{otherwise.} \end{cases} \tag{3}$$

The Huber-norm $\|\cdot\|_\varepsilon$ allows the regulariser to constrain the gradient of disparity to a L^2 norm within a range ε and out of that range a L^1 norm forming a total variation (TV) model so that ε can adjust the degree of undesired staircasing effect and is normally set to 0.01 [16]. The effect of the regulariser is adjusted by λ. To design an image-driven anisotropic regulariser which can maintain disparity discontinuity across image edges, the function w is defined as:

$$w(\mathbf{x}) = \exp(-\alpha \|\nabla I(\mathbf{x})\|_2). \tag{4}$$

Specifically, where a region has high edge magnitude, the output of this weighting function becomes low, which reduces the effect of the regulariser. We can flexibly adjust the support of the exponential function by setting variable α.

Since Eq. 2 is non-convex in the data term and only convex in the regulariser term, to discover the global minimum, conventional approaches for optical flow or variational reconstruction algorithm resort to coarse-to-fine scheme [16]. This requires a good initial state for the global minimum to be found. In addition, reconstruction of coarser layers can lose details in the scene. By contrast, having a cost-volume helps us to avoid the expensive warping scheme. Following a recent large displacement optical flow algorithm [13], we decouple the data term and regulariser term by an auxiliary function $\mathbf{a} : \boldsymbol{\Omega}_I \rightarrow \mathcal{D}$ to form a new energy functional:

$$E(\mathbf{u}, \mathbf{a}) = \int_{\boldsymbol{\Omega}_I} \Big\{ w(\mathbf{x}) \|\nabla \mathbf{u}(\mathbf{x})\|_\varepsilon + Q(\mathbf{u}(\mathbf{x}), \mathbf{a}(\mathbf{x})) + \lambda C(\mathbf{x}, \mathbf{a}(\mathbf{x})) \Big\} \, \mathrm{d}\mathbf{x}, \tag{5}$$

where

$$Q(\mathbf{u}(\mathbf{x}), \mathbf{a}(\mathbf{x})) = \frac{1}{2\theta} (\mathbf{u}(\mathbf{x}) - \mathbf{a}(\mathbf{x}))^2. \tag{6}$$

The first part $w(\mathbf{x})\|\nabla \mathbf{u}(\mathbf{x})\|_\varepsilon + Q(\mathbf{u}(\mathbf{x}), \mathbf{a}(\mathbf{x}))$ is actually a Huber-L^2 model [1] which is similar to TV-L^2 Rudin-Osher-Fatemi model [11] and its global minimum can be found by using an efficient primal-dual algorithm [1,4] for solving the $\mathbf{u}$. Given a temporary solution $\mathbf{u}$, the global minimum of the later part $Q(\mathbf{u}(\mathbf{x}), \mathbf{a}(\mathbf{x})) + \lambda C(\mathbf{x}, \mathbf{a}(\mathbf{x}))$ can be simply found by performing an exhaustive search on $\mathbf{a}(\mathbf{x})$ among the disparities in the cost-volume. θ should be set as a small number to ensure $\mathbf{a}(\mathbf{x}) \simeq \mathbf{u}(\mathbf{x})$ when the algorithm converges.

The primal-dual algorithm works in continuous space so can directly achieve sub-pixel accuracy. Furthermore, the rate of convergence is $O(n)$ [1] which means we need only a few iterations to finish the process. This yields a very efficient and effective convex optimisation in contrast to a traditional global method such as the graph cuts. Following the cone model example in Fig. 1, Fig. 2 shows the convergence curves of the Huber-L^1 convex optimisation using different measures. ZNCC requires much fewer iterations to converge. The reconstruction result is

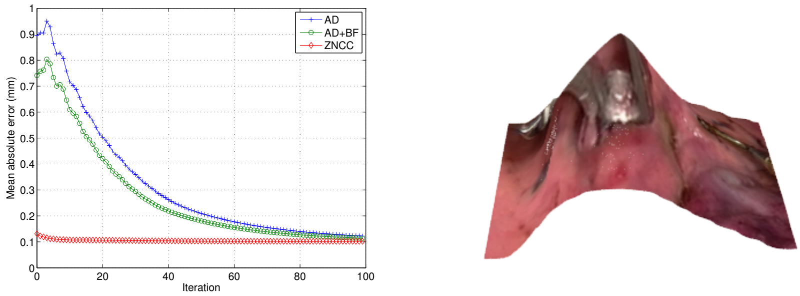

Fig. 2: Left: Convergence of the primal-dual algorithm using different measures as data term. The mean absolute error is calculated by comparing the reconstructed depth with the ground truth depth. Right: The result after the Huber-L^1 convex optimisation.

also shown in Fig. 2. One can observe the point cloud is now much smoother and the staircasing effect due to the discrete cost-volume has been largely eradicated.

3 Empirical Studies

All experiments are conducted on a workstation equipped with 3.1 GHz quad cores CPU and one NVIDIA GeForce GTX 670 graphics card with 2 GB global memory. To maximally exploit the power of parallel computation, all the calculations including the cost-volume construction and the convex optimisation are implemented in CUDA. Currently the proposed reconstruction approach is able to run at 20 fps with the resolution $|\mathbf{\Omega}_I| = 360 \times 288$ and $|\mathcal{D}| = 32$.

We first conduct a noise study to evaluate the robustness for different measures. The proposed approach is then quantitatively evaluated using a cardiac phantom dataset with an independent ground truth. Images in real robot-assisted laparoscopic prostatectomy are reconstructed for qualitative evaluation. In all experiments, only the disparity range $\mathcal{D}$ is dynamic and the rest of parameters for the convex optimisation are set as constants $\{\epsilon, \alpha, \theta, \lambda\} = \{0.01, 0.5, 0.1, 50\}$. The convex optimisation is finished in 150 iterations or if the energy function appears to have converged.

3.1 Noise study

To investigate the robustness of different data terms, we intentionally add white noise to the stereo images of the cone model shown in Fig. 1a. In this experiment the disparity range is set as $\mathcal{D} = [50, 80]$. The resulting reconstruction mean absolute errors (MAE) under different noise variance are reported in Table 1. The results show that there is not much difference between different measures when the image is clean. However, when the noise level becomes large, the measure using simplest pixel-to-pixel AD degrades significantly. In contrast, AD+BF and ZNCC, which perform local cost aggregation, remain accurate in the presence of noise. ZNCC has the best performance in all cases.

Table 1: Under different degrees of noise σ, the reconstruction MAE (mm) compared with the ground truth after the convex optimisation using different data-fidelity terms for the stereo pair of the cone model.

	$\sigma = 0$	$\sigma = 0.01$	$\sigma = 0.015$	$\sigma = 0.02$
AD	0.121	0.623	0.877	2.035
AD+BF	0.121	0.189	0.798	1.521
ZNCC	**0.102**	**0.185**	**0.661**	**1.487**

3.2 Cardiac phantom experiment

The proposed algorithm is quantitatively evaluated by two cardiac datasets collected from [14] which have an associated registered CT model as ground truth as shown in Fig. 3. It should be noted that the ground truth is generated by a 3D/2D point-based registration algorithm, which will inevitably introduce some errors.

Before doing the reconstruction, the stereo image pair are rectified by the provided camera calibration. We further remove the black background by setting an intensity threshold, since such a background does not occur in real surgical images and also it may cause bias when comparing different algorithms. The disparity images are cropped by 15 pixels at the image borders when doing the statistics. In this experiment the disparity range is set as $\mathcal{D} = [0, 30]$.

In Table 2, the MAE and root mean square error (RMSE) to the ground truth point are reported for different real-time dense algorithms using a single stereo pair. The corresponding standard deviation among all frames is also reported. The reconstruction results for a single frame are shown in Fig. 3.

Structure propagation using sparse feature points (SPFP) [15] is a real-time quasi-dense method and fast cost-volume filtering (FCVF) [8] is a local edge-preserving filtering method. A recent real-time dense reconstruction using temporal information (DRTI) algorithm [10] that produces highly accurate reconstruction is also compared, and we compare results of MAE with the best results quoted in their paper. It is evident that our algorithm outperforms the others.

Table 2: Statistics of different algorithms with respect to MAE, RMSE and the percentage of reconstructed points compared with the ground truth.

		Proposed Approach	SPFP [15]	FCVF [8]	DRTI [10]
Cardiac1	MAE(mm)	1.24 ± 0.89	2.36 ± 0.92	4.87 ± 0.87	1.45
	RMSE(mm)	1.85 ± 0.82	3.876 ± 0.87	8.24 ± 0.92	N/A
	Density(%)	100	92	100	N/A
Cardiac2	MAE(mm)	1.47 ± 1.23	3.20 ± 1.15	5.37 ± 1.53	1.53
	RMSE(mm)	2.658 ± 1.47	4.85 ± 1.82	7.73 ± 1.56	N/A
	Density(%)	100	90	100	N/A

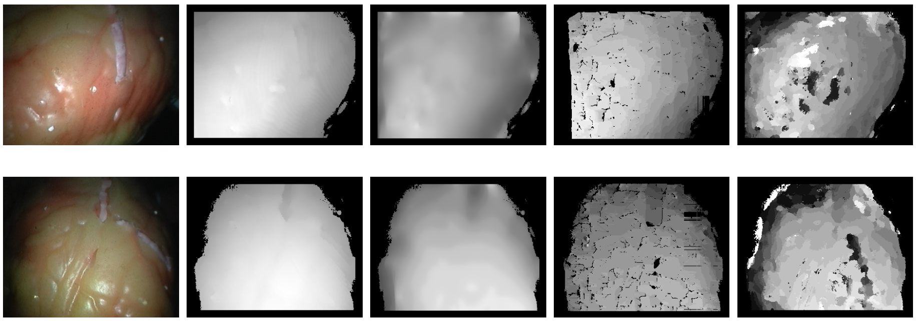

Fig. 3: The cardiac phantoms datasets. The disparity maps showing the reconstruction results from left to right: Ground truth, the proposed approach, SPFP [15] and FCVF [8]

3.3 Qualitative evaluation in *in vivo* images

To qualitatively evaluate the performance of the proposed approach on *in vivo* images, endoscopic stereo images from real robot-assisted laparoscopic prostatectomy are reconstructed as shown in Fig. 4 and in an accompanying video[1]. The overall geometry is well captured. Specular highlights may still cause some mis-matching, which can be resolved by fusing temporal models, and we will investigate this idea as part of our future work.

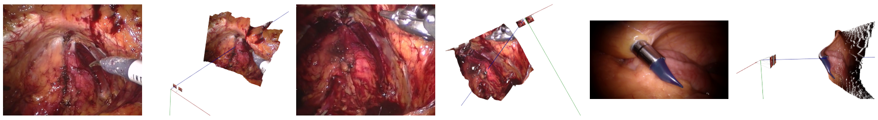

Fig. 4: Qualitative evaluation of the reconstruction results using the proposed approach. The images are obtained from stereo endoscopic camera in real robot-assisted surgery. We recommend to view these images on-screen and zoomed in.

4 Conclusions

In this paper, we have proposed an efficient and effective dense stereo reconstruction method using convex optimisation with a cost-volume. Empirical studies have shown that our reconstruction results outperform the current state of the art methods for endoscopic images and can also run in real-time on the GPU. This is a significant advancement towards improved vision-based tracking of the endoscope and is an important step towards providing image guidance to endoscopic procedures. In our future work, we will be developing dense camera tracking techniques and will extend the current algorithm to fuse a sequence of video images. This will improve the reconstructed model and provide more advanced means for tackling the occlusion at instrument-tissue boundaries.

[1] http://www.doc.ic.ac.uk/~pc3509

References

1. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision 40(1), 150–145 (2011) 4
2. Chang, P., Chen, D., Cohen, D., Edwards, P.: 2D/3D registration of a preoperative model with endoscopic video using colour-consistency. In: Augmented Environments for Computer-Assisted Interventions (AE-CAI) in Conjunction with MICCAI. vol. 7264, pp. 1–12 (2012) 1
3. Devernay, F., Mourgues, F., Coste-Maniere, E.: Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery. In: International Workshop on Medical Imaging and Augmented Reality. pp. 16–20 (2001) 1, 2
4. Handa, A., Newcombe, R.A., Angeli, A., Davison, A.J.: Applications of legendre-fenchel transformation to computer vision problems. Tech. Rep. DTR11-7, Department of Computing at Imperial College London (2011) 4
5. Lau, W.W., Ramey, N.A., Corso, J.J., Thakor, N.V., Hager, G.D.: Stereo-based endoscopic tracking of cardiac surface deformation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). vol. 3217, pp. 494–501 (2004) 2
6. Mirota, D.J., Ishii, M., Hager, G.D.: Vision-based navigation in image-guided interventions. Annual Review of Biomedical Engineering 13, 297–319 (2011) 1, 2
7. Mountney, P., Stoyanov, D., Yang, G.Z.: Three-dimensional tissue deformation recovery and tracking. IEEE Signal Processsing Magazine 27(4), 14–24 (2010) 1, 2
8. Rhemann, C., Hosni, A., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3017–3024 (2011) 3, 6, 7
9. Richa, R., Bo, A.P.L., Poignet, P.: Towards robust 3D visual tracking for motion compensation in beating heart surgery. Medical Image Analysis 15(3), 302–315 (2011) 2
10. Rohl, S., Bodenstedt, S., Suwelack, S., Dillmann, R., Speidel, S., Kenngott, H., Muller-Stich, B.P.: Dense GPU-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. Medical physics 39(3), 1632–45 (2012) 2, 3, 6
11. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1-4), 259–268 (1992) 4
12. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision (1), 131–140 (2002) 2
13. Steinbrucker, F., Pock, T., Cremers, D.: Large displacement optical flow computation without warping. In: IEEE International Conference on Computer Vision (ICCV). pp. 1609–1614 (2009) 4
14. Stoyanov, D.: Stereoscopic scene flow for robotic assisted minimally invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). vol. 7510, pp. 479–86 (2012) 6
15. Stoyanov, D., Scarzanella, M., Pratt, P., Yang, G.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). vol. 6361, pp. 275–282 (2010) 1, 2, 6, 7
16. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic Huber-L1 Optical Flow. In: British Machine Vision Conference (BMVC). pp. 108.1–108.11 (2009) 4