2D/3D Registration of a Preoperative Model with Endoscopic Video using Colour-consistency

Ping-Lin Chang[1], Dongbin Chen[2], Daniel Cohen[2] and Philip "Eddie" Edwards[2]

[1] Department of Computing
Imperial College London, United Kingdom
p.chang10@imperial.ac.uk
[2] Department of Surgery and Cancer
Imperial College London, United Kingdom
Eddie.Edwards@imperial.ac.uk

Abstract Image-guided surgery needs an effective and efficient registration between 2D video images of the surgical scene and a preoperative model of a patient from 3D MRI or CT scans. Such an alignment process is difficult due to the lack of robustly trackable features on the operative surface as well as tissue deformation, and specularity. In this paper, we propose a novel approach to perform the registration using PTAM camera tracking and colour-consistency. PTAM provides a set of video images with the corresponding camera positions. Registration of the 3D model to the video images can then be achieved by maximization of colour-consistency between all 2D pixels corresponding to a given 3D surface point. An improved algorithm for calculation of visible surface points is provided. It is hoped that PTAM camera tracking using a reduced set of points can be combined with colour-consistency to provide a robust registration. A ground truth simulation test bed has been developed for validating the proposed algorithm, and empirical studies have shown that the approach is feasible, with ground truth simulation data providing a capture range in ±9mm/$^\diamond$ with a TRE less than 2mm. Our intended application is robot-assisted laparoscopic prostatectomy.

1 Introduction

Minimally invasive surgery (MIS) is an increasingly popular treatment option due to reduced operative trauma compared to traditional open surgery, and can provide benefits of lower expense, shorter recovery, and reduced incidence of post-surgical complications. In order to perform an operation through small incisions in the skin, MIS uses endoscopic devices to indirectly observe the surgical scene. Due to the nature of live endoscopic video, however, there are severe constraints on the surgeon's spatial perception and reduced operative visual information. In the example of robot-assisted laparoscopic prostatectomy (RALP), though the da VinciTM system provides a magnified 3D visualization along with intuitive scaled manual interaction, the rates of complication from this procedure are still comparable to open surgery. In this scenario, an auxiliary system providing additional visual information would be advantageous.

With the aim of improving outcomes in MIS, image guidance using augmented reality (AR) is proposed. Specifically, by registering a preoperative 3D model to the corresponding 2D endoscopic view of the patient, surgeons can properly orient themselves with respect to the anatomy, which can result in a safer, more effective and potentially more efficient surgery.

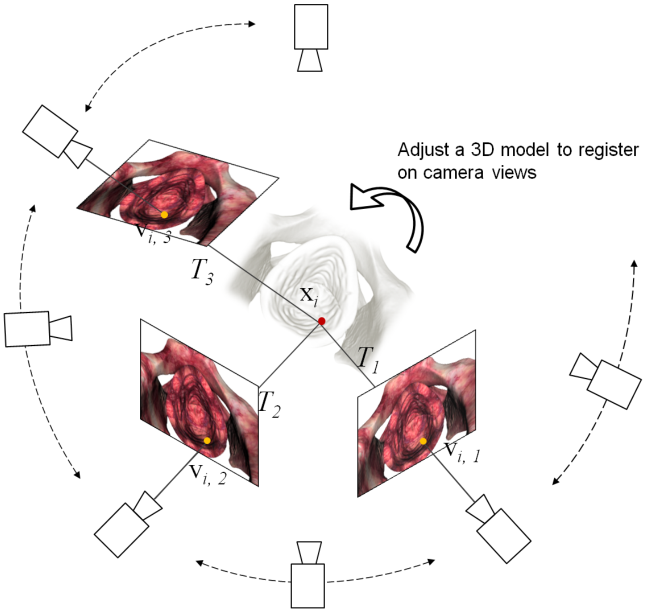
Image registration has been widely studied in computer vision and medical imaging for decades and there have been significant achievements in some areas. However, the registration between a 3D model and 2D endoscopic images remains a difficult problem due to intraoperative tissue deformation, a lack of clear surface features, and the effects of severe specularity. While intraoperative 3D tomography techniques offer precise information about soft tissue morphology and structure, they introduce significant challenges in instrument design, image quality and cost [6]. On the other hand, insufficient features in the endoscopic images make directly reconstructing a deformable 3D surface difficult [15]. Though stereo surface reconstruction is possible [17], errors due to the small baseline between stereo-endoscopic cameras coupled with the small visible region in the endoscopic view mean that these surfaces may not be suitable for registration.

In this paper, we propose to combine parallel tracking and mapping (PTAM) [5] and a colour version of colour-consistency [2] to carry out the 3D to 2D registration. Unlike feature-based registration algorithms, a pixel-based colour-consistency approach may be more robust when insufficient features are visible in the scenes. The role of PTAM in our method is to provide correct camera pose. By optimizing a similarity measure, the 3D model can then be adjusted to a pose which is the most consistent among all the camera views.

The lack of a ground truth in 3D medical image registration has led to the suggestion of simulations for algorithm testing [4], We have developed a ground truth simulation to validate the performance of the proposed registration algorithm. Additionally, a fast method for calculation of the visible 3D surface points for colour-consistency calculation is proposed. Finally, we propose a registration approach for endoscopic surgery of the lower abdomen.

2 Related works

A significant issue when reconstructing a 3D organ is tissue deformation. This is not directly tackled in this paper. Algorithms for deformable 3D surface reconstruction can be separated into template-based and non-rigid structure from motion reconstruction. Both approaches have shown success in deformable 3D surface reconstruction [15]. However, when there are too few features that can be detected in the scene, neither class of approaches perform well, which prevents them from being used in practice. Nevertheless, a number of techniques have been published which applied feature-based 3D reconstruction in endoscopic sequences. Stoyanov et al presented a method for dense depth recovery from stereo laparoscopic images of deformable soft-issue [16]. Mourgues et al proposed a correlation-based stereo method for surface reconstruction and organ modelling from stereo endoscopic images [11]. Quartucci Forster et al applied a shape-from-shading technique to estimate the surface shape by recovering the depth information from the surface illumination [14]. Mountney et al proposed a probabilistic framework for selecting the most discriminative descriptors by Bayesian fusion method to compare twenty-one different descriptors [9]. Wang et al used scale-invariant feature transform (SIFT) features for endoscopy sequences and used the adaptive scale kernel consensus for robust motion estimation [18]. Wu et al. also tracked SIFT features and utilized an iterative factorization method for structure estimation [19]. Mountney et al presented a technique to construct a 3D map of the scene for MIS endoscopic surgery while recovering the camera motion based on simultaneous localization and mapping (SLAM) from a stereo endoscope, but their main focus was surface reconstruction [10]. SLAM and parallel tracking and mapping (PTAM) [5] have achieved significant success in real-time camera tracking and mapping for real scenes. Newcombe and Davison also utilized PTAM for dense real-time surface reconstruction [12]. However, with the features available in endoscope images, the performance of 3D reconstruction or registration can be expected to be worse.



(a)

Figure 1. Colour-consistency with PTAM's camera tracking for 3D to 2D image registration.

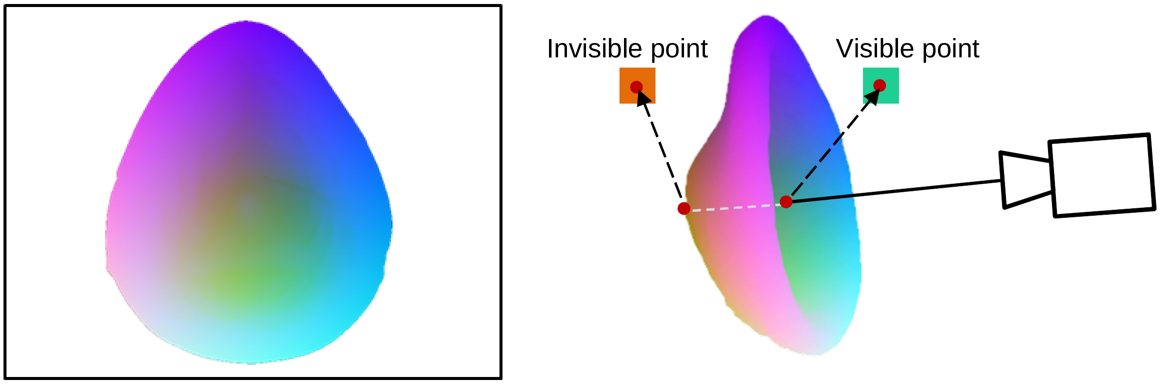
3 PTAM and colour-consistency

Details for PTAM's design can be found in [5]. Multiple robust features are tracked in the scene and the camera tracking and scene reconstruction are calculated in separate parallel threads. In our work, the main role of PTAM is to provide the camera poses for a number of video frames. One could argue that the points from the map that PTAM creates could be used for registration. We argue that this is not a good strategy in our case as there will be many points that do not lie on our preoperative surface that are still useful for camera tracking. Also, the use of a pixel intensity-based method should give a denser and more robust match than one based on relatively sparse features. Note that PTAM does not provide the scale but we propose that for the da Vinci$^{\mathrm{TM}}$, which incorporates a stereo endoscope, we can overcome the scaling issue using stereo. The idea of incorporating colour-consistency with PTAM's camera tracking for 3D to 2D image registration is shown in Fig. 1.

First, it is necessary to determine which model vertices are visible in which keyframes. The visibility can be checked by projecting each 3D vertex, $\mathrm{x}_i = (x_i, y_i, z_i)^{\mathrm{T}}$, onto a 2D pixel, $\mathrm{v}_{i,n} = (u_{i,n}, v_{i,n})^{\mathrm{T}}$, using:

$$s\mathrm{v}_{i,n} = \mathrm{KT}_n\mathrm{x}_i, \tag{1}$$

where $\mathrm{T}_n = [\mathrm{R}_n|\mathbf{t}_n]$ is a rigid camera transformation for keyframe n with a rotation R_n and a translation $\mathbf{t}_n$, K is the camera intrinsic matrix and s is a scalar [3]. The cost function is the average of the variance of colour in each vertex as follows:



(a)

(b)

Figure 2. (a) A camera frame as the front view of a 3D bladder model mapped by a RGB model. (b) The proposed fast visible point detection algorithm. Both cyan (front) and orange (back) points are projected onto the same position in the camera frame, and can be distinguished by the vertex colour.

$$C = \frac{1}{N}\sum_{\mathrm{x}=0}^{N}\{\frac{1}{3n_{\mathrm{x}}}\sum_{i=0}^{n_{\mathrm{x}}}[(r_{i,\mathrm{x}} - \bar{r_{\mathrm{x}}})^2 + (g_{i,\mathrm{x}} - \bar{g_{\mathrm{x}}})^2 + (b_{i,\mathrm{x}} - \bar{b_{\mathrm{x}}})^2]\} \tag{2}$$

where N is the total number of vertices xs that are visible in at least two keyframes, n_{x} is the number of visible frames for a vertex x, $r_{i,\mathrm{x}}$, $g_{i,\mathrm{x}}$ and $b_{i,\mathrm{x}}$ are the RGB colour components and $\bar{r_{\mathrm{x}}}$, $\bar{g_{\mathrm{x}}}$ and $\bar{b_{\mathrm{x}}}$ are the mean RGB value for a vertex x.

Lighting and reflectance of the surface are significant factors which will affect colour consistency. For preliminary studies, we use purely ambient lighting so that we have an ideal environment to examine the proposed approach. Only vertices visible in at least two keyframes are taken into account. For each point on our surface, we first need to calculate whether it is visible in each keyframe. To achieve this we set a surface colour for each vertex, where the colour $(r, g, b)^{\mathrm{T}}$ has been set to the position $(x, y, z)^{\mathrm{T}}$ as shown in Fig. 2 (a). By rendering the object from each keyframe position we can limit ourselves to the visible front face simply by checking that the projected colour matches the colour of the vertex. As shown in Fig. 2 (b), the colour of the vertex at the front side is different from the back side. This proves to be more efficient than z-buffer methods which require us to calculate the distance to the vertex in each of the images.

4 Empirical studies

The lack of a ground truth in 3D medical image registration has led to the suggestion of simulations for algorithm testing [4][8]. We use a simulated phantom video with realistic texture for the bladder and pelvis. The origianl phantom model is shown in Fig. 3 (a). The bladder and pelvis models are manually segmentaed from a CT scan as shown in Fig. 3 (b). We then render a 3D model with a real surgical scene as texture as shown in Fig. 3 (c). This can give us the ground truth model pose and the true camera positions and orientations when a simulation video is generated. To generate the simulation video, we

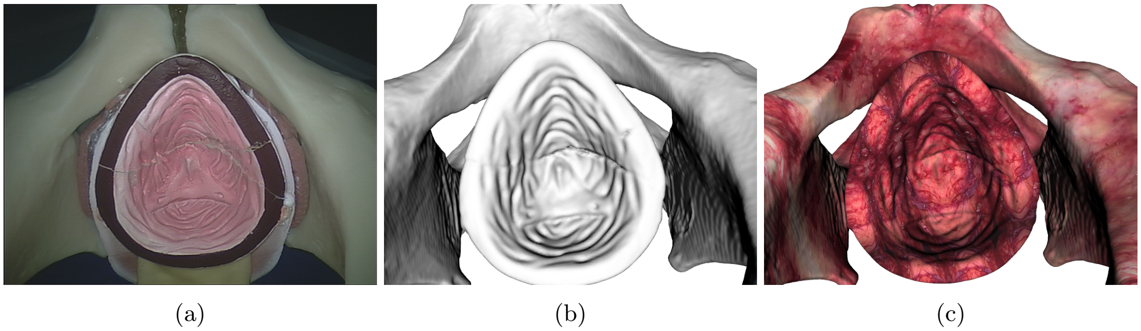


Figure 3. The original phantom model (a), the bladder and pelvis models manually segmentaed from a CT scan (b) and a real surgical scene textured on them (c).

have implemented a program using Qt and VTK taking the intrinsic and extrinsic parameters of the real stereo endoscope equipped on da Vinci^{TM} into account. The real stereo endoscope was calibrated by using Bouguet's camera calibration toolbox [1]. Note that currently the lighting condition is set to full ambient lighting with no surface diffusion and specular reflection. PTAM can then be tested on this simulation video. Furthermore, given the estimated camera positions and rotations of the set of keyframes captured from PTAM, the colour-consistency algorithm can be investigated.

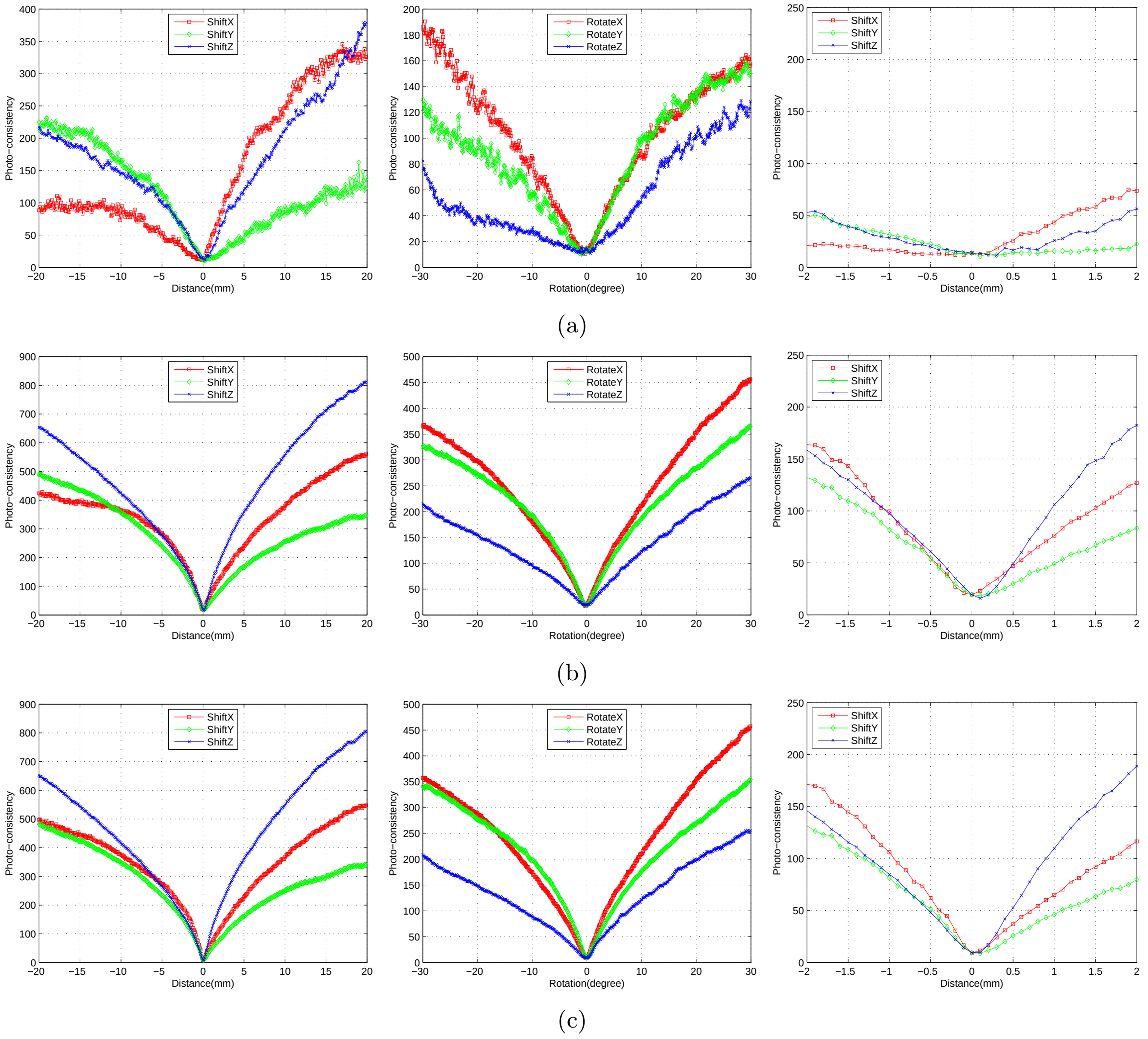
The proposed approach is to combine PTAM [5] and a colour version of photo-consistency [2] to carry out the 2D to 3D registration. The role of PTAM in our method is to provide correct camera pose. By optimizing a similarity measure, the 3D model can then be adjusted to a pose which is the most consistent among all the camera views.

Currently we use the original implementation of PTAM by Klein and Murray to extract the evaluated camera positions and rotation matrices [5]. Experiments were run on an Intel(R) Core(TM) 2 Quad 2.5 GHz CPU with 4GB physical memory and a nVidia GeForce GT 330 graphic card. All programs are implemented by C++ and CUDA C. With GPU programming, the visible points detection process takes about 2.32ms and the calculation of the colour-consistency takes about 2.23ms with 61,856 vertices in the 3D model and two 768x576 pixels keyframes.

4.1 Gold standard evaluation

The task is to optimize the transformation matrix of the 3D model to accurately align to the 2D projection imagrs. For a 3D rigid transformation matrix, the minimum parameterlization is three for rotation and three for translation. Observing the colour-consistency space with the ground truth camera poses may help us investigating the problem. We devised a test to intentially move each parameter apart from the ground truth pose in -30° to 30° in 0.1° interval and -20mm to 20mm in 0.1mm interval for rotation and translation respectively. Fig. 4 (a) and (b) reveals that the cost function has clear global minimum in the investigated range as well as an advantage when more keyframes obtained from PTAM are used. With more keyframes involved, the cost functions should become smoother which results in a more robust optimization.

Projected vertices are assigned colours by using nearest neighbour scheme. Another option is to apply bilinear interpolation to calculate approximate colours for projected vertices according to a four-neighbour-pixel relationship. Fig. 4(c) shows that although the bilinear interpolation scheme can produce smaller values of colour-consistency and have more precise values at the ground truth pose, the overall curve shape is very similar to the one using nearest neighbour scheme.



(a)

(b)

(c)

Figure 4. The validation results produced by rotating and shifting the 3D model with ground truth camera poses using 2 keyframes (a), using 20 keyframes (b), using 20 keyframes with bilinear interpolation scheme(c). The charts at the most right hand side are enlarged version of shift tests.

4.2 Optimization with the model pose

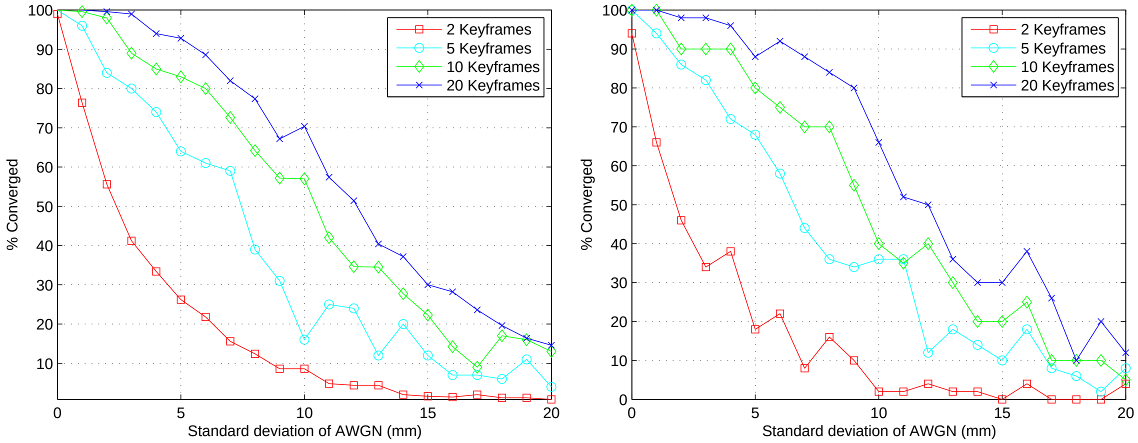
In initial study we assumed accurate camera poses are provided. To optimize the pose of the 3D model, a local optimizer is employed to optimize the rigid body transformation matrix $\mathrm{T}_{model} = [\mathrm{R}_{model}|\mathbf{t}_{model}]$ which has six degree of freedom. A derivative-free algorithm is more preferable since we have no gradient information for the cost function. Although using finite-difference to evaluate the Jacobian matrix for non-linear gradient-based algorithms such as Newton, Gauss-Newton and Levenberg-Marquardt is possible, the calculation is highly expensive since every parameter requires a colour-consistency evaluation in each iteration. Having tried classical derivative-free approaches, we found Powell's BOBYQA [13] has better performance than Principal Axis (PRAXIS) and Simplex. In addition, BOBYQA provides bound-constrained optimization which can restrict the search within a reasonable capture range.

We conducted an experiment which changes the ground truth pose by using additive white Gaussian noise (AWGN) with different standard deviations. Under each standard deviation, we ran 500 times for the registration using a random set of 2, 5, 10 and 20 keyframes. A three-layers pyramid suggested by [7] for the derivative-free optimizer was used. If the target registration error (TRE) which is defined as root mean squared error (RMSE) of the entire vertices is less than 2mm, we regarded the registration process has converged.

Fig. 5 (a) and (b) show the results of the optimization using the nearest neighbour and using the bilinear interpolation respectively. The performance is not much difference between them, and curves in each case share common trends. One can see when only 2 keyframes are used, the frequency of convergence drastically decreases at the very beginning. When using 20 keyframes, the convergence rate starts to drop down after 3mm standard deviation which is about the range in ±9mm/°. Involving more keyframes may result in even better performance, but it also introduces more computation effort. Tab. 1 shows the statistics of the required optimizing iteration and the running time for the converged cases. Note that the iteration time is accumulated by the number of iteration the optimizer takes in each layer. The average iteration times are almost consistent no matter how many keyframes are used. This is because the required iteration of a derivative-free optimizer is only affected by the number of parameters, and here we have 6 parameters in all cases.

Table 1. Average running time and required iteration time for optimization with only the model pose

Keyframe #	Avg. iteration	Avg. running time per iter. (sec.)	Avg. running time (sec.)
2	130	0.05	6.75
5	149	0.12	14.32
10	151	0.20	30.48
20	154	0.40	62.06



(a)

(b)

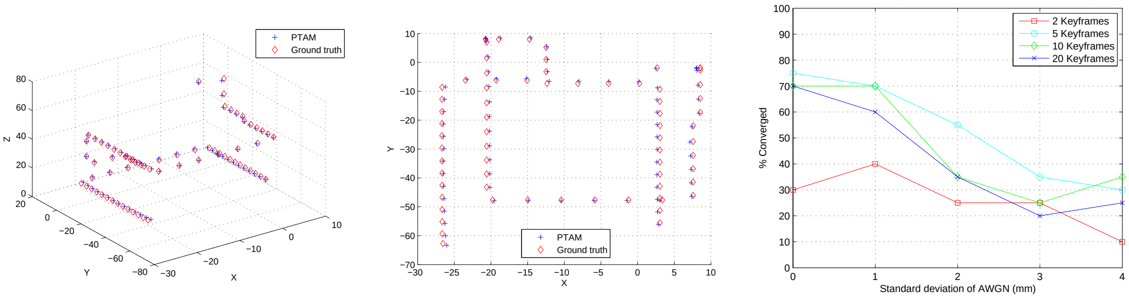
Figure 5. The frequency of convergence of the proposed approach using nearest neighbour (a) and using bilnear interpolation (b) in the optimization of only the model pose.

4.3 Optimization with the camera poses

PTAM tracks camera poses on the fly by simultaneously tracking features and mapping scenes. To validate the feasibility of using PTAM's camera poses, we run PTAM on the ground truth simulation video to obtain the tracked camera positions and rotation matrices, and then, such estimated camera poses are compared with the ground truth camera poses. Since PTAM's coordinate system is defined by using a stereo initialization, to fairly compare them, we use a rigid registration, Procrustes analysis algorithm, to carry out transforming coordinate systems.

Fig. 6 shows one of the results after we transform PTAM's camera positions into the ground truth's coordinate system with a sum of squared error 0.5 mm on average. As can be seen although the two sets of points are fairly close to each other, there are small errors in each corresponding pair. These errors including camera's positions and orientations will propagate to the registration result of colour-consistency.

The camera poses therefore need to be optimized. When the optimization involves all cameras poses, there are $N \times 6 + 6$ where N is the number of cameras and each pose has 6 DoF. Using the derivative-free optimizer, BOBYQA, and the pervious experiment setting, the considerable number of parameters makes the optimization prone to fail as shown in Fig. 6 (c). Note that using 20 keyframes is no longer giving better performance since there are 126 parameters to be optimized. Instead, the less number, 5 keyframes, has a better frequency of convergence. Tab. 2 shows the average iteration times also have a considerable increase due to the increased number of parameters. This proves that a more efficient and effective optimization algorithm is necessary to our approach.



(a) (b) (c)

Figure 6. The result of Procrustes analysis transforming PTAM's camera positions into the ground truth's coordinate system (a). The vertical view of X-Y plane (b). Note that the scale is millimeter. The frequency of convergence of the proposed approach using nearest neighbour in the optimization of the model and camera poses (c).

Table 2. Average running time and required iteration time for optimization with the model and camera poses

Keyframe #	Avg. iteration	Avg. running time per iter. (sec.)	Avg. running time (sec.)
2	383	0.08	30.64
5	580	0.21	121.80
10	690	0.43	296.70
20	953	0.72	686.16

5 Conclusions

We have presented a novel approach to registration of a preoperative 3D model to intraoperative endoscopic video which combines PTAM tracking with colour-consistency registration, which incorporates a fast calculation of the visible 3D surface. To validate the method we developed a simulation test bed with accurate ground truth. This could be used to validate other reconstruction or registration algorithms.

Much work remains to be done. The mapping side of PTAM is in some senses unnecessary as we have our preoperative map. Also, the errors in PTAM's camera tracking must be optimized as well, which together with the optimization of the 3D model pose can be related to bundle adjustment problem [3].

We will use our simulation test bed to establish the robustness of the method under different levels of noise, blurring and specular reflection. We need to establish the accuracy of our approach in the real clinical setting. The aim will be to match a 3D model of the pubic arch from preoperative imaging to the laparoscopic view during robot-assisted prostatectomy. We are in the process of gathering pre- and intraoperative clinical data for this purpose.

Acknowledgements

We would like to thank Cancer Research UK for funding this research under project C24520/A8087. We are also grateful to the radiology and theatre staff at Imperial College Healthcare NHS Trust for their ongoing assistance and to other members of the Department of Computing VIP group in for advice and guidance.

References

1. J.-Y. Bouguet, "Complete camera calibration toolbox for Matlab®." [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/ 5
2. M. Clarkson, D. Rueckert, D. Hill, and D. Hawkes, "Using photo-consistency to register 2D optical images of the human face to a 3D surface model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1266–1280, 2001. 2, 5
3. R. I. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed. Cambridge University Press, 2004. 3, 9
4. P. Jannin, J. M. Fitzpatrick, D. J. Hawkes, X. Pennec, R. Shahidi, and M. W. Vannier, "Validation of medical image processing in image-guided therapy," *IEEE transactions on medical imaging*, vol. 21, no. 12, pp. 1445–1449, Dec. 2002. 2, 4
5. G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *The 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 2007, pp. 225–234. 2, 3, 5
6. S. Lavallee, P. Sautot, J. Troccaz, P. Cinquin, and P. Merloz, "Computer-assisted spine surgery: A technique for accurate transpedicular screw fixation using CT data and a 3-D optical localizer," *Journal of Image Guided Surgery*, 1995. 2
7. F. Maes, D. Vandermeulen, and P. Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information." *Medical image analysis*, vol. 3, no. 4, pp. 373–86, Dec. 1999. 7
8. P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš, "A review of 3D/2D registration methods for image-guided interventions," *Medical image analysis*, Apr. 2010. 4
9. P. Mountney, B. Lo, S. Thiemjarus, D. Stoyanov, and G.-Z. Yang, "A probabilistic framework for tracking deformable soft tissue in minimally invasive surgery," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2007)*, 2007, vol. 4792, pp. 34–41. 2
10. P. Mountney, D. Stoyanov, A. J. Davison, and G.-Z. Yang, "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2006*, vol. 4190, 2006, pp. 347–354. 2
11. F. Mourgues, F. Devernay, and E. Coste-Maniere, "3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery," in *IEEE and ACM International Symposium on Augmented Reality - ISAR 2001*, 2001, p. 191. 2
12. R. A. Newcombe and A. J. Davison, "Live dense reconstruction with a single moving camera," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, 2010. 2
13. M. J. D. Powell, "The bobyqa algorithm for bound constrained optimization without derivatives," *Technical Report: Department of Applied Mathematics and Theoretical Physics, Cambridge England*, 2009. 7
14. C. H. Quartucci Forster and C. L. Tozzi, "Towards 3D reconstruction of endoscope images using shape from shading," in *Brazilian Symposium on Computer Graphics and Image Processing*, vol. 0, 2000, p. 90. 2
15. M. Salzmann and P. Fua, "Deformable surface 3D reconstruction from monocular images," *Synthesis Lectures on Computer Vision*, vol. 2, no. 1, pp. 1–113, 2010. 2
16. D. Stoyanov, A. Darzi, and G.-Z. Yang, "Dense 3D depth recovery for soft tissue deformation during robotically assisted laparoscopic surgery," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2004)*, 2004, vol. 3217, pp. 41–48. 2
17. D. Stoyanov, M. 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 2010)*, 2010, vol. 6361, pp. 275–282. 2
18. H. Wang, D. Mirota, M. Ishii, and G. Hager, "Robust motion estimation and structure recovery from endoscopic image sequences with an adaptive scale kernel consensus estimator," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 2008, pp. 1–7. 2
19. C.-H. Wu, Y.-N. Sun, Y.-C. Chen, and C.-C. Chang, "Endoscopic feature tracking and scale-invariant estimation of soft-tissue structures," *IEICE - Transactions on Information and Systems*, vol. E91-D, pp. 351–360, 2008. 2