Detection and Localization of Instruments in Minimally Invasive Surgery

Max Allan*, Sébastien Ourselin†, Steve Thompson†, David J. Hawkes†, John Kelly‡, Danail Stoyanov*

*Centre for Medical Image Computing and Dept. of Computer Science, UCL

†Centre for Medical Image Computing and Dept. of Medical Physics and Bioengineering, UCL

‡Division of Surgery and Interventional Science, UCL Medical School
{maximilian.allan.11,s.ourselin,s.thompson,j.d.kelly,danail.stoyanov}@ucl.ac.uk

***Abstract*—Integration of vision techniques for detection and localization of instruments in Minimally Invasive Surgery (MIS) can provide assistance to surgeons through motion control and guidance aids. In this study, we propose a framework for detecting and localizing the tool pose in 3D in the coordinate system of the observing camera. Using Random Forest classifiers to probabilistically label pixels in laparoscopic images as instrument or tissue, we recover the pose in 3D through a region based level-set segmentation technique. We demonstrate the effectiveness of the proposed system by comparison to ground truth tracking data obtained through an optical tracking system. This work is a continuation of the work in [1].**

I. Introduction

Minimally Invasive Surgery (MIS) has greatly improved surgical outcomes compared to open surgery by reducing the trauma of the procedure by accessing the surgical site through small keyhole ports [2]. However difficulties arising from the removal of direct access to the underlying anatomy has complicated surgery. Precise control of the instruments and tissue is more difficult to achieve with minimally invasive instruments and perceptive information such as force feedback and field of view is reduced. Robotic and computer assisted surgery have improved the surgeon's ability to navigate and operate [3], [4] through the introduction of HD stereoscopic video cameras and motion control systems that reduce hand tremors and remove the motion inversion of controlling a pivoting instrument.

As MIS is introduced to more complex procedures, the benefits of integrating preoperative information and navigation assistance become more prominent. Real-time knowledge of the instrument position and orientation with respect to both the surgical camera and the underlying anatomy becomes increasingly important as a means of introducing these systems. Computer vision techniques have demonstrated themselves to be effective methods of obtaining this information with minimal modification to existing surgical workflows.

There have been numerous proposals in the literature to adapt computer vision techniques to the minimally invasive environment. Several methods have attempted to integrate low level image features as part of point-based pose estimation methods [5], [6] and have achieved successful results on surgical images. Template tracking has also been attempted [7], [8]. However, one of the major shortcomings of this type of technique is that small scale features are vulnerable to occlusions so often found in surgical environments due to tissue, blood and smoke.

An alternative approach which makes use of region statistics when localizing pose presents a solution to this problem. [9] suggested an approach that involves aligning a model with a class probability map and achieve robust results. However, the method suffers from a non-differentiable cost function which leads to slow and unpredictable optimizations. Approaches which make use of level-set based region alignments have been successful in the computer vision literature [10], [11] and it is the approach of [12] which we follow most closely. The motivation behind following these methods is that the level set function can be smoothly differentiated with respect to the pose parameters of the target object leading to minimization techniques that simultaneously find the optimal 3D pose.

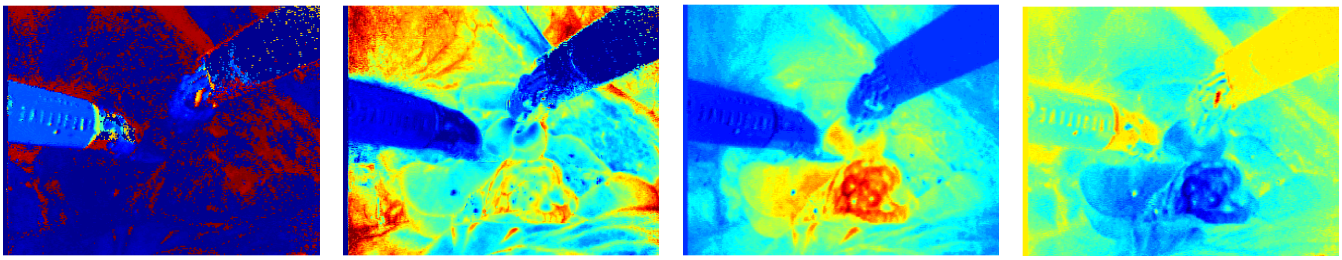
In this work we propose a 3 stage method of localizing medical instruments in 3D given a simple shape and appearance model. We perform detection with a random forest classifier after extensive experimentation to select suitable features, a 2D initialization with a shape analysis approach and 3D pose refinement with a level-set based segmentation.

II. Method

A. Detection

Random Forests (RF) [13] were chosen as the method of discriminating the regions of the image which correspond to the instrument and the tissue. They provide an accurate, fast and potentially parallelizable method which can be easily extended to handle multi-class training data, a useful feature when classifying multiple distinct tool or tissue types.

1) Feature selection: The complexity of light reflectance in the scene leads to some visual ambiguities between the instrument and areas of the tissue surface which makes classification challenging. As such, feature selection for the random forest becomes extremely important to ensure that the classification is accurate enough to enable good pose estimation.



(a) Hue (b) Saturation (c) Opponent 2 (d) Opponent 3

Fig. 1: Visualisations of the discriminative power of the chosen colourspaces.

We tested several different features across distinct regions of MIS images: lightly colored tool shafts, darkly colored tool shafts, instrument tips and tissue. We performed our tests on several color features: RGB, HSV, CIE XYZ, Opponent 2 and Opponent 3, further details on these color spaces are reported in [14], [15].

2) Forest Implementation: We made use of the Random Forest implementation of the OpenCV library [1] and limit our forest size to 50 trees of no more than 10 levels to increase speed of both training and classification.

B. Instrument Pose Detection

1) Parametrization: We choose the standard parametrization of the instrument using the Euler angles (θ, ψ, ϕ) and a translation $\mathbf{t}$ in the camera coordinate system. As we have so far only modelled the instrument shaft as a featureless cylinder (without considering articulation of the instrument head) we are unable to detect the axial rotation parameter ϕ leaving our model with 5 recoverable degrees of freedom.

2) 2D Initialization: As our 3D pose estimation technique relies on a gradient descent based method to refine the pose, it is important to obtain a good initial estimate to avoid local minima and to reduce the time taken to converge. We achieve this by selecting the largest connected regions of the classification map and then finding their principal axes using the moment of inertia tensor. Then by assuming that the connected region can be approximated by a 2D cylinder, we can estimate its width and height (in pixels) from the eigenvalues of the tensor as $r = \sqrt{2I_2/m}$ and $l = \sqrt{12I_1/m - 3r^2}$. The estimate of width can be used to estimate the depth of the instrument given that we know the width of the instrument in metric units a priori. From this initialization we obtain an estimate of $\mathbf{t}$ as well as the rotational parameter θ.

3) 3D Refinement: To estimate the additional rotational degree of freedom ψ as well as refine the estimates of θ and $\mathbf{t}$ we employ the 3D pose estimation technique of [12]. Given an estimate of image pose and a known 3D model of the target object, a segmentation can be defined through the outer contour of the its projection onto the image plane. The quality of the segmentation can be defined using an image functional:

$$E(g(\mathbf{p})) = -\sum_{\mathbf{p}\in\Omega} \Big(\log \left(H_e(g(\mathbf{p}))P_f + (1 - H_e(g(\mathbf{p}))P_b \right) \Big) \tag{1}$$

where $P_{f|b}$ refers to the confidence of a pixel belonging to the instrument class and the tissue class respectively and is computed by the random forest as the fraction of trees which voted for each class. $g(\mathbf{p})$ represents the level set function and, as usual, is represented by a signed distance function (negative values for the exterior, positive for the interior). $H(.)$ is a smoothed Heaviside function which filters the distance values into interior/exterior regions with the smoothing allowing for uncertainty in the location of the contour. To find the minimum of Eq. 1 from the set of possible projections of the target object we search the space of its pose parameters with gradient descent.

III. Results

The results were collected by a C++ implementation of the algorithm without any parallelization or optimization of the code and as such is not real-time. Random Forest mean classification time is 3.47 seconds for an SD image and pose estimation takes 1-20 seconds, with the large variation due to differences in accuracy of the starting estimate.

A. Features for Classification

We performed several experiments to determine the most discriminative visual features to use for classification. We used the general statistical technique for measuring the distance between histograms of the Bhattacharyya distance [16] and also an internal estimator provided by all decision tree methods known as variable importance [17]. To ensure we considered a range of tissue types and instrument appearances we perform tests on 97 manually labelled images from 6 different surgical procedures.

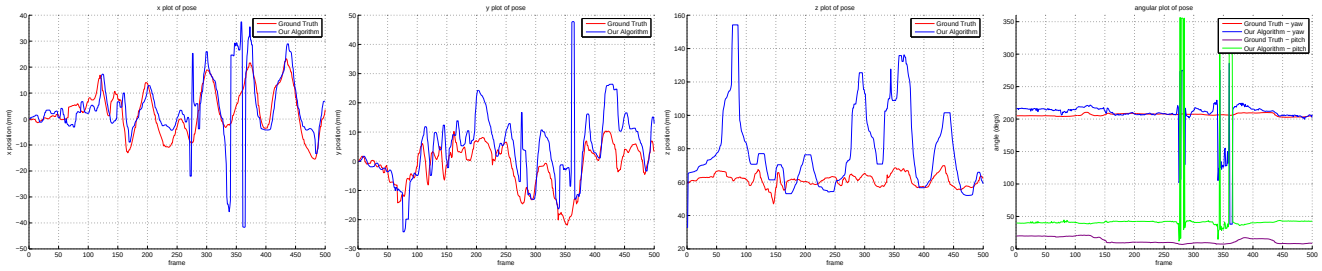
Our investigations showed that the most discriminative colour spaces appear to be the hue, saturation and the opponent 2 and 3 colour spaces.

B. Laboratory experiments

Validation of the pose estimation was achieved with an experiment where data was acquired from an NDI Optotrak Certus[2] optical tracking system. We manufactured rigid bod-

[1] www.opencv.willowgarage.com

[2] http://www.ndigital.com/lifesciences/certus-motioncapturesystem.php



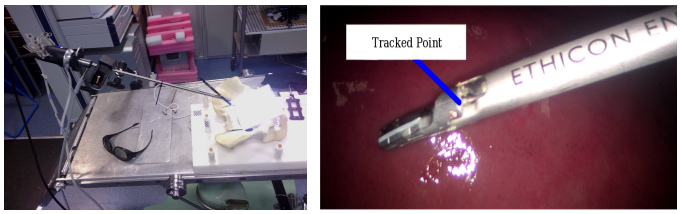
(a) (b) (c) (d)

Fig. 2: The error plots for the Optotrak calibrated pose. As can be seen the errors are quite low, particularly in x and y. The z error occasionally moves far from the true estimate (which in turn distorts the x and y estimates). The positional estimates have been median filtered to smooth out some of the errors.

ies with embedded optical markers and attached one to the proximal end of a Viking Systems 3DHD laparoscope and one to the proximal end of an Ethicon monopolar dissector. The experimental setup can be seen in Figure 3. Camera and hand eye calibration between the camera and Optotrak coordinate systems were obtained using toolboxes available online to allow comparison of measurements made by our method using the camera and those from the Optotrak [3] [4]. Camera calibration reprojection error was 0.2 pixels and additionally, the instrument was calibrated to measure the offset between the tool's tip and the attached rigid body with a calibration error of 0.17mm $\pm$ 0.18mm.

Experiments were performed by translating and rotating the instrument between the camera and an *ex vivo* lamb liver tissue sample while the Optotrak system recorded the instrument pose in the camera coordinate system for each frame. An instrument appearance model was learned from images of the tool in front of a homogeneous background and a background appearance model was learned from separate images of the liver.

By computing the 3D pose of the instrument at each frame and comparing to data from the Optotrak system we show motion plots of the tip position and error plots in Figure 2. To compensate for calibration error, which results in a constant offset, we show the motion after the coordinate systems have been matched for the first frame. The plots visibly show that our method correctly localises to the ground truth pose in the majority of frames. The mean and standard deviation of the error for the instrument are 0.171mm and 0.182mm whereas for the camera they are 0.981mm and 0.002mm respectively. These are promising results considering we are estimating 3D information from a monocular image and this explains the larger error visible in the z axis. Occasional errors are also visible as spikes in our measurements, particularly around frames 280 and frame 350. These occur when the instrument is moved out of the view of the camera and the pose localisation system incorrectly recognises part of the shadow in the image as an instrument. These errors could easily be removed by incorporating a simple tracking framework such as Kalman filtering.



(a) (b)

Fig. 3: (a) shows the laboratory experimental setup with the laparoscope complete with optical tracking markers. (b) shows a frame from the laparoscope showing the instrument in front of a lamb's liver. The marker indicates the point on the instrument which is tracked by our system.

IV. Conclusion

In this study we have proposed a method for detecting and localizing instruments in surgical images. Experimentation with optically calibrated data shows that our method is robust and generally accurate with some inaccuracy in the depth estimation. As typical surgical systems are equipped with stereo cameras our future work will include integrating stereo constraints to improve accuracy. We further hope to make use of temporal tracking estimates to reduce the error from noisy estimates and additionally we plan to integrate extensive GPU parallelization to achieve real time performance of our work, as has been reported in the literature.

References

[1] M. Allan, S. Ourselin, S. Thompson, D. J. Hawkes, J. Kelly, and D. Stoyanov, "Toward detection and localization of instruments in minimally invasive surgery," *IEEE transactions on bio-medical engineering*, vol. 60, pp. 1050–1058, Apr. 2013. PMID: 23192482.

[2] A. Darzi and S. Mackay, "Recent advances in minimal access surgery," *BMJ*, vol. 324, pp. 31–34, Jan. 2002.

[3] D. J. Mirota, M. Ishii, and G. D. Hager, "Vision-based navigation in image-guided interventions," *Annual Review of Biomedical Engineering*, vol. 13, pp. 297–319, Aug. 2011. PMID: 21568713.

[3] http://www0.cs.ucl.ac.uk/staff/Dan.Stoyanov/calib/

[4] http://www.vision.ee.ethz.ch/software/calibration_toolbox//calibration_toolbox.php

[4] D. Stoyanov, "Surgical vision," *Annals of Biomedical Engineering*, vol. 40, pp. 332–345, Feb. 2012.

[5] S. Voros, J. Long, and P. Cinquin, "Automatic detection of instruments in laparoscopic images: A first step towards high-level command of robotic endoscopic holders," *The International Journal of Robotics Research*, vol. 26, pp. 1173 –1190, Nov. 2007.

[6] A. Reiter, P. K. Allen, and T. Zhao, "Learning features on robotic surgical tools," in *Computer Vision and Pattern Recognition*, 2012.

[7] R. Sznitman, K. Ali, R. Richa, R. Taylor, G. Hager, and P. Fua, "Data-driven visual tracking in retinal microsurgery," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2012* (N. Ayache, H. Delingette, P. Golland, and K. Mori, eds.), vol. 7511 of *Lecture Notes in Computer Science*, pp. 568–575, Springer Berlin / Heidelberg, 2012.

[8] D. Burschka, J. J. Corso, M. Dewan, W. Lau, M. Li, H. Lin, P. Marayong, N. Ramey, G. D. Hager, B. Hoffman, D. Larkin, and C. Hasser, "Navigating inner space: 3-D assistance for minimally invasive surgery," in *In: Workshop Advances in Robot Vision, in conjunction with the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 67–78, 2004.

[9] Z. Pezzementi, S. Voros, and G. D. Hager, "Articulated object tracking by rendering consistent appearance parts," in *IEEE International Conference on Robotics and Automation, 2009. ICRA '09*, pp. 3940 –3947, May 2009.

[10] S. Dambreville, R. Sandhu, A. Yezzi, and A. Tannenbaum, "Robust 3D pose estimation and efficient 2D region-based segmentation from a 3D shape prior," in *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, (Berlin, Heidelberg), pp. 169–182, Springer-Verlag, 2008.

[11] B. Rosenhahn, T. Brox, and J. Weickert, "Three-dimensional shape knowledge for joint image segmentation and pose estimation," in *Pattern Recognition, volume 3663 of LNCS*, p. 109116, Springer, 2005.

[12] V. A. Prisacariu and I. D. Reid, "PWP3D: Real-Time segmentation and tracking of 3D objects," *International Journal of Computer Vision*, vol. 98, pp. 335–354, Jan. 2012.

[13] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[14] G. W. Meyer and D. P. Greenberg, "Perceptual color spaces for computer graphics," in *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '80, (New York, NY, USA), p. 254261, ACM, 1980.

[15] T. Gevers and H. Stokman, "Classifying color edges in video into shadow-geometry, highlight, or material transitions," *Multimedia, IEEE Transactions on*, vol. 5, pp. 237 – 243, June 2003.

[16] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[17] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognition*, vol. 44, pp. 330–349, Feb. 2011.