Image Based Surgical Instrument Pose Estimation with Multi-class Labelling and Optical Flow

Max Allan[1], Ping-Lin Chang[1], Sébastien Ourselin[1], David J. Hawkes[1], Ashwin Sridhar[2], John Kelly[2], and Danail Stoyanov[1]

[1] Centre for Medical Image Computing, University College London, UK
[2] Division of Surgery and Interventional Science, UCL Medical School, UK

Abstract. Image based detection, tracking and pose estimation of surgical instruments in minimally invasive surgery has a number of potential applications for computer assisted interventions. Recent developments in the field have resulted in advanced techniques for 2D instrument detection in laparoscopic images, however, full 3D pose estimation remains a challenging and unsolved problem. In this paper, we present a novel method for estimating the 3D pose of robotic instruments, including axial rotation, by fusing information from large homogeneous regions and local optical flow features. We demonstrate the accuracy and robustness of this approach on ex vivo data with calibrated ground truth given by surgical robot kinematics which we will also make available to the community. Qualitative validation on in vivo data from robotic assisted prostatectomy further demonstrates that the technique can function in clinical scenarios.

1 Introduction

Robotic minimally invasive surgery can facilitate procedures in confined and difficult to access anatomical regions. However, accessing the anatomy with robotic instruments reduces the surgeon's ability to sense force feedback from instrument-tissue interactions and the limited field of view of the surgical camera makes localization with respect to preoperative patient data challenging. Computer assisted interventions (CAI) can integrate additional information during the operation to help the surgeon and knowing the 3D position and orientation of the surgical instruments during surgery is a critical CAI element. The instrument pose can additionally be used in robotic surgery to provide control enhancements with dynamic motion constraints or to detect tool-tissue interactions and provide force feedback [13].

Image-based methods can potentially estimate instrument pose in the reference frame of the laparoscope without requiring electromagnetic or optical sensors [6,12]. This usually involves extracting image features such as edges, points or regions and then solving alignment cost functions which measure the agreement with parametrized models of the tool [10]. Gradient based methods

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 331–338, 2015.
DOI: 10.1007/978-3-319-24553-9_41

are often preferred but it is challenging to develop cost functions that do not easily become trapped in local minima and fail to find the correct pose [15,1]. [9] used gradient free optimization from color and texture features for articulated instruments but the chosen cost can be complex to optimize resulting in slow and often inaccurate solutions. Another alternative is to use Random Forests (RF) to detect instrument parts [14] which gives promising results and low computational cost but is only shown as a 2D tracking method. Using robot kinematic information from the joint encoders has been investigated but accumulation of errors can result in significant error and bounded brute-force template-matching has been employed to reduce the offset [3]. Region based methods for surgical instruments were proposed in [1] where bag-of-pixel based object appearance models were used to demonstrated pose estimation that is robust to viewpoint and illumination changes [4]. However disregarding all spatial information within the object boundary makes it challenging to recover the instrument roll axis and the yaw axis which is usually strongly affected by the foreshortening visual cue. Additional cues have been fused with region features to obtain more stable tracking but did not address the correspondence problem when dealing with multiple point detections on the instrument tip [2].

In this paper, we present a novel image-driven pose estimation technique for robotic instruments in minimally invasive surgery (MIS). This is achieved by fusing large scale region based constraints with low level optical flow information. The interior homogeneous-intensity regions of the instruments are described with separate appearance models and this is used to formulate region based alignment as a multi region problem rather than using a binary silhouette. The interior instrument appearance is a strong regional cue on robotic instruments and helps to solve the foreshortening problem by introducing a full visible boundary in the image plane. We focus on estimating rigid 3D pose without the full articulation of the robotic instruments. Quantitative validation is shown on calibrated ex vivo data collected using the da Vinci® research kit (DVRK) and API [8] and qualitative validation is demonstrated on challenging in vivo data.

2 Method

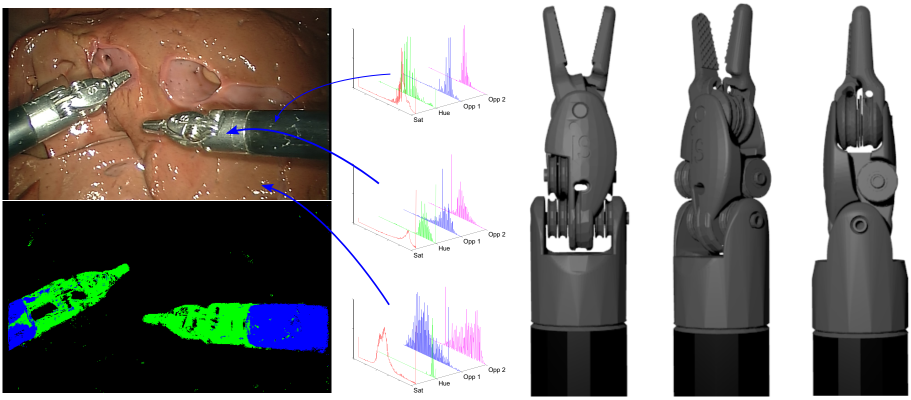
Our method works by fusing large-scale region features, which are based on the output of multi-label probabilistic classification, with small-scale flow features. The region features drive coarse pose estimation through the alignment of predicted regions generated from the projection of the instrument, given a particular pose estimate, with the detected regions on the classification map. To improve fine scale estimation, salient features on the instrument surface are tracked from frame-to-frame using optical flow.

2.1 Multi-label Probabilistic Classification

We use RFs to provide probabilistic region classification an image, assigning pixel to one of K object classes, where in a typical image there will be $K - 1$

regions for an instrument and 1 region for the background (see Fig. 1a. When applied to classification, each RF is an ensemble of decision trees which each vote on a labelling for the input pixel. The vote of a single tree is decided by directing an input sample $\mathbf{x}$ from a root node $\S_{parent}$ to one of its two child nodes $\S_{child}$ according to a linear model $y = \mathbf{w}\mathbf{x}$ where the left child is chosen if $y < T_i$ and to the right node if $y \geq T_i$ where T_i is a node specific threshold value. This root to child splitting is applied recursively on the sample until it reaches a terminating node, known as a *leaf node* where it is given a label according to a probability distribution $p(C|\mathbf{x})$ stored in that node. Each tree in the forest applies a classification vote and then these votes are averaged across all the trees to obtain the output of the forest.

Training our forest involved the manual segmentation of a single frame containing instruments positioned in front of a tissue background into the specified K classes however, in principal a large background library of possible tissue types and foreground models would be learned offline to allow the system to operate in different surgical setups without re-training. We use a simple color based feature set of Hue, Saturation, Opponent 1 and Opponent 2 which were shown to have good classification on MIS images [1].



(a)

(b)

Fig. 1. (a) shows the feature distribution for each of the $K = 3$ classes with output classification and (b) shows example renderings of a robotic instrument CAD model from Intuitive Surgical Inc.

2.2 Multi-region Segmentation with Level Sets

Statistical region-based 3D pose estimation is formulated as a segmentation problem, where the pose parameters which enable the silhouette of the projection of a geometric model to optimally divide an image into two regions are estimated. For robotic instruments, modelling internal homogeneous regions separately can be used to create strong delineating contours, which can improve the estimated pose over modelling the interior with a single distribution and using only the silhouette. We therefore model an instrument's appearance with $K-1$ statistical models so, given a single background model, we describe the image with K statistical models. Pose estimation then becomes a problem of finding the $K-1$ contours which divide the image plane up into K regions such that the pixels within the i^{th} region agree maximally with the i^{th} statistical model.

We describe the segmenting contours using level sets of signed distance functions [7,4] as they avoid the problem of an explicitly parametrized curve while elegantly applying implicit correspondences between the region-based data contours and the model projection contours. Finding the contours which optimally assign each of the K models to the image becomes a variational problem which is described using the the following cost:

$$E_{region}(\Theta) = -\sum_{i}^{K}\sum_{\mathbf{x}\in\Omega} \log\left(H(\phi^i(\mathbf{x},\Theta))P_f^{\Omega_i} + (1-H(\phi^i(\mathbf{x},\Theta)))P_b^{\Omega_i}\right) \qquad (1)$$

where $\phi^i(\mathbf{x},\Theta)$ is the Euclidean distance between at the pixel $\mathbf{x}$ and the closest point on the contour generated from the i^{th} model projection at pose Θ. $\phi(.)$ is set to the negative distance outside the contour and positive inside. We represent pose as translation and rotation for which we use the quaternion representation. $H(.)$ is a smoothed Heaviside function which truncates the values of ϕ into a spatial prior on the model assignment. Ω_i are the pixels within the i^{th} region of the image Ω and $P_{f,b}^{\Omega_i}$ are the learned distributions for the pixels inside (foreground) and outside (background) the i^{th} contour. Rather than performing one-against-all for the background distribution, we instead use the expected neighbour class of the pixel $\mathbf{x}$ as the chosen background distribution.

2.3 Optimization and Tracking

The level set segmentation provides accurate pose estimation but in the presence of fast motion and noise errors can appear especially around the roll axis of the instrument where the contour cues may not provide sufficient constraints. However, low level interior features and optical flow in the image can provide strong cues about the motion of the instrument from frame-to-frame. We use simple gradient-based salient features [11] and assuming the first frame contains the correct pose, backproject the tracked points onto the object model and do frame-to-frame tracking using the Lucas-Kanade method [5]. The optimization of this functional involves the joint minimization of the region based cost which we perform over both frames, if available, and the flow based cost solving for a single set of pose parameters to obtain stereo constraints.

$$E(\Theta) = \min_{\Theta} \sum_j E^j_{region}(\Theta) + \lambda \sum_i^N ||\mathbf{y}'_i - P(\mathbf{y}_i, \Theta)||^2 \tag{2}$$

where $\mathbf{y}'_i$ is the position of the i^{th} optical flow tracked point in the image frame and $P(\mathbf{y}_i, \Theta)$ represents our estimate of where the point $\mathbf{y}_i$ projects to from the surface of our model. In essence this is 2D-3D registration. The sum over j is over the left and right frames of the stereo pair where E^j_{region} refers to the energy function from the left or right frame. λ is the usual weighting factor between our two cost functions and is set experimentally. We use gradient descent to find the minimum for each frame and combine with a Kalman filter for temporal consistency. Initialization is assumed to be correct for the first frame and can be achieved within a few seconds using a manual positioning of the instrument.

3 Results

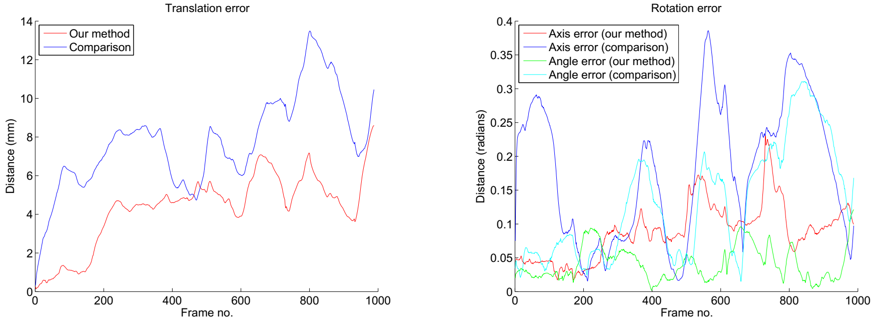
Our algorithm is written in C++ and OpenGL using OpenCV[1]. Processing time measured on a single core of a 1.9GHz processor for classification of a single stereo frame using a RF is $\approx$ 0.83 seconds, for a gradient descent step on one stereo frame is $\approx$ 0.3 seconds (typically 10-20 steps required) and processing time for the flow tracking is $\approx$ 0.006 seconds per frame. The most computationally expensive component is the region based cost for which each pixel is computed independently allowing for real time speeds when using a GPU implementation [10]. Furthermore, RFs are suitable for GPU parallelisation and by only performing classification in regions where the derivatives are non-zero, we can greatly reduce the number of pixels which require classification to around 0.5% of the image. The source code and data from our method are available online[2].

3.1 Quantitative Validation

Using the da Vinci® API it is possible to estimate the position and orientation of each robotic instrument by reading the motor joint encoder values and then using the Denavit Hartenberg (DH) chain to compute the relative orientations of the instruments and the camera. The encoder values accumulate errors over time resulting in joints being offset from the camera frame, which is why image-based estimation is important even when encoder information is available. The majority of this error can be calibrated out with a fixed offset to obtain a ground truth pose in the camera frame with high accuracy. We constructed an ex vivo sequence of 1000 frames with a lamb liver tissue and computed the pose of the instruments in each frame using our method and also computed a ground truth pose using the robot forward kinematics. To assess our method, we compare

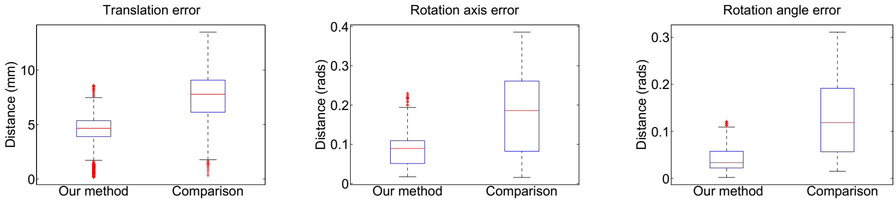
[1] http://opencv.org

[2] http://www.surgicalvision.cs.ucl.ac.uk/code



(a) **(b)**

Fig. 2. (a) RMS error in measuring translation from the camera center to the origin of the model coordinates (near the head) from our algorithm and from the comparison method when compared to the ground truth estimates using ex vivo data. (b) Similar to (a) but showing angular distance between axis and angular error when using the angle-axis representation of the instrument pose.



(a) **(b)** **(c)**

Fig. 3. Error distribution for data in Fig. 2. Red line shows median error while the top and bottom of the box show the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

it with a our previous silhouette based tracking technique [2] which does not use interior contours or the low level flow features.

We report errors for rotation using the angle-axis representation, computing angular distance between the axes and also the difference between the angular rotation around those axes, and translation between the camera and instrument coordinate system, where we average the error at each frame across both instruments. Trajectory errors are shown in Fig. 2 and error distributions are show in a box plot in Fig. 3. Translational error is broken down into each axis in numerical form in Table 1.

Table 1. The mean error ± std deviation of each translational degree of freedom and the rotational angle/axis of the robotic instruments for the ex vivo data. Top row is our method and bottom is the comparison method.

	x (mm)	y (mm)	z (mm)	axis (rads)	angle (rads)
This work	0.70 ± 0.31	0.50 ± 0.27	4.09 ± 1.82	0.08 ± 0.04	0.04 ± 0.03
[2]	1.09 ± 0.65	0.59 ± 0.29	7.48 ± 2.32	0.18 ± 0.10	0.13 ± 0.08

3.2 Qualitative Validation

We show qualitative validation for both ex vivo and in vivo sequences in Fig. 4 where the overlap between the projected model and the underlying image demonstrates the accuracy of our method.

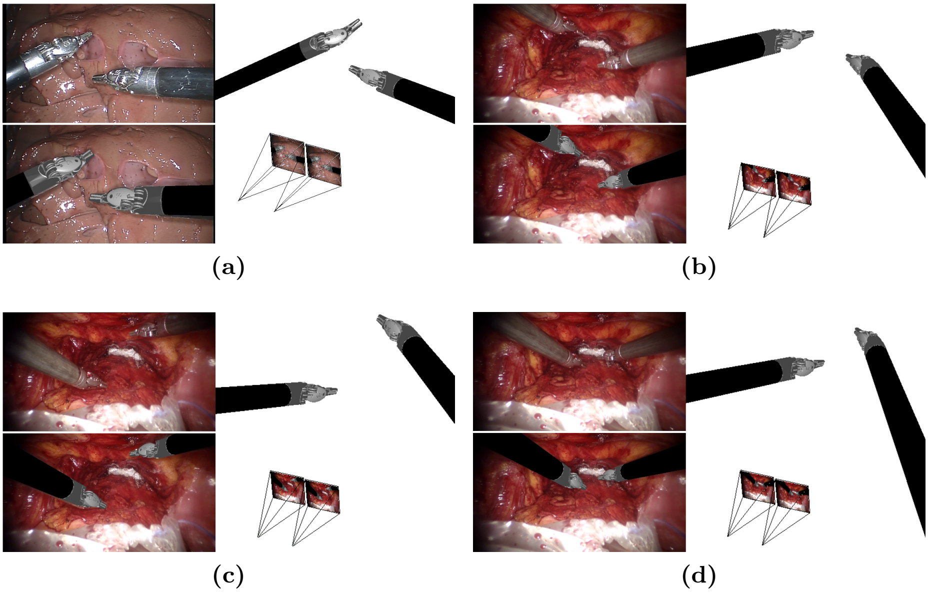


Fig. 4. Qualitative validation on a challenging ex vivo (a) and in vivo (b-d) sequence showing an example left camera image, the same frame with instruments overlaid at the current pose estimate and a 3D plot of the instruments in front of the stereo camera pair.

4 Conclusion and Discussion

The results from Fig. 2 and Table 1 demonstrate the significant quantitative improvements in our algorithm in comparison to the state-of-the-art method [2], in particularly with respect to the rotational parameters, which is the result of using frame-to-frame low level flow features to estimate the instrument roll. Results in the x and y direction are very stable with errors below ≤ 0.7 mm with increased error in the z direction. This is typically the largest source of error as, even when using stereo constraints larger differences in 3D position only reveal themselves as small inaccuracies in the 2D data/model alignment. Additionally we observe that the errors increase over the duration of the experiment which is common in model based tracking as errors from previous frames gradually cause the correct estimate to drift away from the true solution. Future improvements to our method will involve solving our cost function for all degrees of freedom of the articulated robotic instrument rather than using $SE(3)$.

Acknowledgements. The authors would like to acknowledge Simon Di-Maio and Dale Bergman at Intuitive Surgical Inc. for supplying the CAD models.

References

1. Allan, M., Ourselin, S., Thompson, S., Hawkes, D.J., Kelly, J., Stoyanov, D.: Toward detection and localization of instruments in minimally invasive surgery. IEEE Transactions on Biomedical Engineering 60(4), 1050–1058 (2013)
2. Allan, M., Thompson, S., Clarkson, M.J., Ourselin, S., Hawkes, D.J., Kelly, J., Stoyanov, D.: 2d-3d pose tracking of rigid instruments in minimally invasive surgery. In: Stoyanov, D., Collins, D.L., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI 2014. LNCS, vol. 8498, pp. 1–10. Springer, Heidelberg (2014)
3. Austin, R.: K, A.P., Tao, Z.: Articulated surgical tool detection using virtually-rendered templates. In: Computer Assisted Radiology and Surgery (2012)
4. Bibby, C., Reid, I.: Robust Real-Time visual tracking using Pixel-Wise posteriors. In: ECCV, pp. 831–844 (2008)
5. Bouguet, J.Y.: Pyramidal implementation of the lucas kanade feature tracker. Intel Corporation, Microprocessor Research Labs (2000)
6. Chmarra, M.K., Grimbergen, C.A., Dankelman, J.: Systems for tracking minimally invasive surgical instruments. Minimally Invasive Therapy & Allied Technologies 16(6), 328–340 (2007)
7. Cremers, D., Rousson, M., Deriche, R.: A review of statistical approaches to level set segmentation. IJCV 72(2), 195–215 (2007)
8. DiMaio, S., Hasser, C.: The da vinci research interface (July 2008)
9. Pezzementi, Z., Voros, S., Hager, G.D.: Articulated object tracking by rendering consistent appearance parts. In: ICRA 2009, pp. 3940–3947 (May 2009)
10. Prisacariu, V.A., Reid, I.D.: PWP3D: Real-Time segmentation and tracking of 3D objects. Int. J. Computer Vision 98(3), 335–354 (2012)
11. Shi, J., Tomasi, C.: Good features to track. In: CVPR 1994, pp. 593–600 (June 1994)
12. Speidel, S., Sudra, G., Senemaud, J., Drentschew, M., Müller-Stich, B.P., Gutt, C., Dillmann, R.: Recognition of risk situations based on endoscopic instrument tracking and knowledge based situation modeling. In: Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling, vol. 6918 (2008)
13. Stoyanov, D.: Surgical vision. Annals of Biomedical Engineering 40(2) (2012)
14. Sznitman, R., Becker, C., Fua, P.: Fast part-based classification for instrument detection in minimally invasive surgery. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part II. LNCS, vol. 8674, pp. 692–699. Springer, Heidelberg (2014)
15. Sznitman, R., Ali, K., Richa, R., Taylor, R.H., Hager, G.D., Fua, P.: Data-driven visual tracking in retinal microsurgery. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 568–575. Springer, Heidelberg (2012)