2D-3D Pose Tracking of Rigid Instruments in Minimally Invasive Surgery

Max Allan[1,2], Steve Thompson[1,3], Matthew J. Clarkson[1,3], Sébastien Ourselin[1,3], David J. Hawkes[1,3], John Kelly[4], and Danail Stoyanov[1,2]

[1] Centre for Medical Image Computing, UCL, London, UK
[2] Department of Computer Science, UCL, London, UK
[3] Department of Medical Physics and Bioengineering, UCL, London, UK
[4] Division of Surgery and Interventional Science, Medical School, UCL, London, UK
{maximilian.allan.11,s.thompson,m.clarkson,s.ourselin,
d.hawkes,j.d.kelly,d.stoyanov}@ucl.ac.uk

Abstract. Instrument localization and tracking is an important challenge for advanced computer assisted techniques in minimally invasive surgery and image-based solutions to instrument localization can provide a non-invasive, low cost solution. In this study, we present a novel algorithm capable of recovering the 3D pose of laparoscopic surgical instruments combining constraints from a classification algorithm, multiple point features, stereo views (when available) and a linear motion model to robustly track the tool in surgical videos. We demonstrate the improved robustness and performance of our algorithm with optically tracked ground truth and additionally qualitatively demonstrate its performance on *in vivo* images.

1 Introduction

Image-based instrument tracking and localization has important applications in computer assisted interventions (CAI) and in robotic minimally invasive surgery (RMIS). Computing the pose of the instruments is critical for enabling enhanced guidance and navigation where precise knowledge of the sub-surface patient anatomy can assist the surgeon to avoid critical structures and accurately excise tissue. With robotic manipulators, virtual fixtures can be applied if the tools approach delicate regions [1] or alternatively haptic feedback can be used to improve instrument-tissue manipulation [2]. The major challenge with localizing the tools is in developing a system that integrates into the operating room with minimal disruption of the workflow or additional invasion of the patient anatomy. While instrument tracking can be realised by using hardware sensors, encoders or external optical systems, such approaches require extensive hardware integration and still have limitations in accuracy and integration into the operating theatre. A significant advantage of image-based methods [3, 4] is that they recover the tool's position and orientation directly in the surgeon's viewing reference and do not require any additional hardware [5, 6].

For minimally invasive surgery (MIS), instrument detection based purely on images has been investigated for a number years [7]. Recent state-of-the-art methods involve the use of trained classifiers and combine the detection and subsequent tracking of instruments [8, 9]. Such algorithms achieve excellent results but from a single image only the 2D image position of the instrument is recovered. The full 3D position and orientation of the instrument can be recovered using specialized fiducial markers machined onto the instruments, however, this approach is restrictive and it interferes with the hardware making it difficult for general theatre use with arbitrary instruments [10]. Naturally appearing features can potentially also be used to localize the instrument. For example, edge information with gradient direction filtering based on the trocar position has been demonstrated [11]. This constraint can cope with significant image noise but estimating the trocar position can be complex in the presence of insufflation and physiological motion such as breathing and heart rate. Gradient based point features can also be combined with color-based features and classification to track articulated robotic instruments [12] or as part of a brute force matching of rendered tool templates [13]. Such methods can be implemented in real-time with GPU processing but they rely heavily on kinematic data from the robotic system and this therefore limits their application to non-robotic procedures. Additionally, the gradient features are focussed around the tip of the articulated instrument which fails to exploit the large constraint provided by the cylindrical instrument shaft. In [14] we demonstrated the use of this constraint for five degrees of freedom (5 DOF) instrument localization.

In this paper we propose combining constraints from feature points with a region based level set segmentation to develop an instrument localization and tracking framework that is more robust than using either individual technique in isolation. We handle challenging data containing occlusions and large reflections by exploiting strong prior knowledge of the instrument appearance and shape though discriminative classification with a Random Forest (RF) and by applying constraints to the level set propagation. We formulate this within a cost function that is simple to optimize and robust to noise in the image. The addition of multi-view constraints to suit an emerging line of stereo laparoscopes add further information and temporal motion is incorporated with a Kalman filter. We show that these modifications provide improvement over previous work by comparison experiments with *ex vivo* tissue and ground truth tracking provided by an optical system. To further illustrate the effectiveness of our algorithm we include qualitative results from MIS videos.

2 Method

2.1 Region based alignment

Region based tracking methods using level sets are generally framed as the maximization of an energy functional

$$E = \int_{\Omega_f} r_f(I(\mathbf{x}), C))d\Omega + \int_{\Omega_b} r_b(I(\mathbf{x}), C)d\Omega \tag{1}$$

where $r_{f|b}$ represent functions which measure the agreement between the information in the pixels $\mathbf{x}$ of image I within a contour C (the foreground) and outside the contour (the background) with learned statistical models. These agreement functions are summed over the foreground and background regions $\Omega_{f|b}$. Normally this energy functional is maximized by finding the set of pose parameters which define the optimal segmentation of the target image into a foreground and background region.

The significant challenges within region based tracking are selecting a function $r(.)$ to measure the region agreement and choosing the parameters which determine the evolution of the contour. By assuming a weak constraint, which can be relaxed, that we are tracking a rigid object we solve the latter problem by following [15] optimizing in the space of the 6 degrees of freedom of a rigid transformation, constraining the contour to belong to the set of image plane projections of our target object at the current estimate of pose.

Selecting the function $r(.)$ is problematic in MIS as the complex lighting and occlusions lead to ambiguous regions for which simple classification models fail. Following [14] we learn the function $r(.)$ with random decision forests trained on the Hue, Saturation, Opponent 2 and Opponent 3 color spaces, which were demonstrated by the authors to have good performance on MIS images.

2.2 Incorporating stereo constraints

A significant challenge of 3D pose estimation using a monocular camera is the difficulty in estimating the depth of the target object purely from perspective cues [14]. Incorporating stereo constraints is important for creating a system that is capable of reliably estimating 3D information. Practically, stereo acquisition is also more common now with 3D laparoscope systems recently becoming available from a variety of commercial manufacturers [16]. We incorporate stereo as a special case multi-view constraint [15] by constructing the cost function over both images of the stereo pair before solving for the pose in the reference camera coordinate system.

2.3 Refinement with point based tracking

One of the challenges of region based tracking is that it struggles to refine the pose to highly accurate solutions when there are ambiguous contours or noise around the edge of the target object. However, it is good at providing a reasonably close solution to the global maximum.

Point based tracking methods however can provide highly accurate pose estimation but suffer heavily from data association errors, particularly when working with relatively featureless surfaces such as those found on medical instruments. The robustness of region based tracking can be combined with the high precision of point based tracking by jointly optimizing for both features. We avoid the difficulties of data association errors by searching for matches in a small region around expected locations of feature points (as suggested by the current estimated pose of the target object).

This results in our overall discretized energy functional being represented as

$$E = \sum_{i \in \mathcal{I}_{l|r}} \sum_{\mathbf{x} \in \Omega_i} \left(r_f(\mathbf{x}) H(g(\mathbf{x})) + r_b(\mathbf{x})(1 - H(g(\mathbf{x}))) \right) + \lambda \sum_{\mathbf{y} \in \Gamma} |\mathbf{y}' - P(\mathbf{y})|^2 \quad (2)$$

where $\mathbf{y}'$ is a matched feature in the image (we perform feature matching exclusively in the left image for simplicity) and $P(\mathbf{y})$ is the projection of its corresponding 3D point. λ is a weighting parameter used to modify the contribution of the point alignments. $H(.)$ is the smoothed Heaviside function of the level set embedding function $g(\mathbf{x})$, which is represented as a signed distance function as is typical in the level set formulation of image segmentations [17]. $\mathcal{I}_{l|r}$ are set of the left and right images (although this could represent any number of calibrated images) and $\mathbf{x} \in \Omega_i$ refer to the pixels in a single image over which segmentation is performed. Γ is the set of features on the target object which we are attempting to match in the image. In our current implementation we choose SIFT features [18] but any feature with good invariance to lighting and pose changes could be chosen. To build a library of detectable points for a given instrument, we collect target SIFT features from a sample image of the object in which the instrument pose has been manually aligned, backprojecting them to their intersection with the target object to find their object space coordinates.

The cost function is optimized using gradient descent as this only requires first derivatives yielding faster iterations than other optimization techniques. We additionally use the quaternion representation of angular pose which, although requiring normalization at each step, avoids the singularity problems of the Euler angle representation.

2.4 Initialization and tracking

To initialize our pose estimate we follow the method of [14]. Frame by frame tracking is provided with a linear Kalman filter for both position and orientation. Our state vector for the k^{th} estimate is defined as

$$\mathbf{x}_k = (x, y, z, \dot{x}, \dot{y}, \dot{z}, \theta, \psi, \phi) \quad (3)$$

where the terms have their usual meanings. We transform the quaternion rotation representation to Euler angles to allow linearization of the Kalman filter. We update pose using the standard Kalman Filter equations

$$\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + N(0, \mathbf{Q}) \quad (4)$$

$$\mathbf{z}_k = \mathbf{M}\mathbf{x}_k + N(0, \mathbf{R}) \quad (5)$$

where $\mathbf{z}_k$ is the measurement vector, $\mathbf{F}$ is the position-velocity state transition matrix and $\mathbf{M}$ is the identity observation model. Both are corrupted by normally distributed noise of zero mean and variance $\mathbf{Q}, \mathbf{R}$. For more details on the linear Kalman filter, the reader is directed to [19].

3 Results

To evaluate the performance of the proposed method we conducted experiments within a controlled laboratory environment where we were able to obtain ground truth data. For comparison to prior work we compared our results to a recent state-of-the-art method [14]. Qualitative evaluation is also reported for *in vivo* surgical videos.

The implementation of the method used in these results is written in C++ and a single iteration of the gradient descent takes approximately 1 second on a 3.0 GHz dual core CPU. As each pixel of the level set optimization is evaluated independently, the method is highly parallelizable and real time performance has been demonstrated for similar techniques on a GPU [15].

3.1 Laboratory experiments

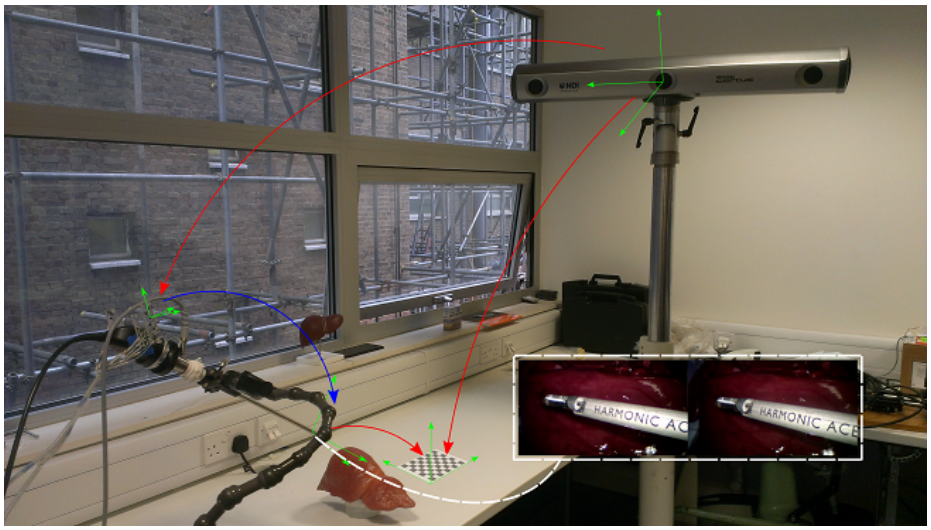
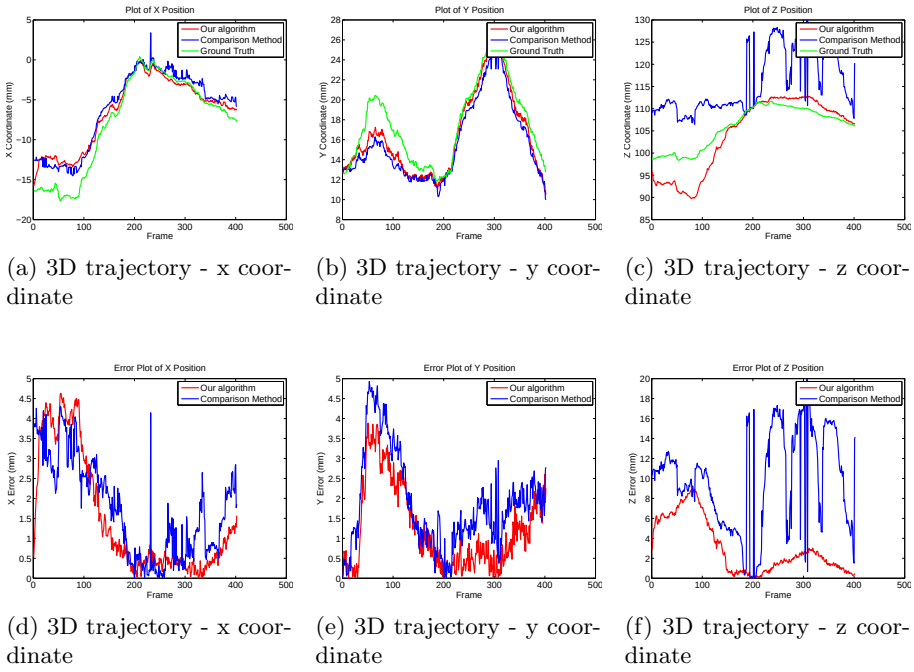


Fig. 1: This image shows the optical tracking system we constructed to capture video with synchronized ground truth data. Inset shows an example frame from our captured video.

A mock-up surgical site was constructed with a lamb's liver and an Endopath monopolar dissector (Ethicon Endo-Surgery Inc.) as the working instrument. The scene was visualized with a 3DHD laparoscope (Viking Systems). We attached optical tracking markers to the proximal end of the laparoscope and to the proximal end of the instrument and tracked their locations using an Optotrak Certus system (Northern Digital). Hand eye calibration was performed using OpenCV[1] and Tsai's handeye method [20] implemented within the NifTK toolbox[2] to determine the transformations between the optical tracker and the

[1] http://docs.opencv.org/
[2] http://cmic.cs.ucl.ac.uk/home/software/

camera coordinate systems (See Figure 1). The location of the instrument tip relative to the tracking markers was found using an invariant point method, also implemented in NifTK. Laparoscope tracking error was experimentally determined to be 1.7mm RMS and instrument tracking error estimated to be 0.7mm RMS, assuming independence this gives a tracking error of 1.8 mm RMS for the instrument tip relative to the laparoscope lens.



(a) 3D trajectory - x coordinate

(b) 3D trajectory - y coordinate

(c) 3D trajectory - z coordinate

(d) 3D trajectory - x coordinate

(e) 3D trajectory - y coordinate

(f) 3D trajectory - z coordinate

Fig. 2: These plots show the ground truth translation from the center of the camera coordinate system to the tip of the tracked instrument obtained with the Optical tracking system compared with the results obtained from our algorithm and the algorithm of [14].

We learn instrument color models from a single image of the target object manually segmented from a homogeneous background and the background model is learned from a single image of the target environment captured before the instrument is introduced to the scene.

We recorded a single video of the instrument moving in front of the liver synchronising the video and tracking data using NifTK. The transformation from the camera coordinate system to the tip of the instrument is computed for each frame by our algorithm and by the optically tracked markers. Due to the calibration inaccuracy we are forced to manually remove the offset by choosing a frame where the tracking alignment appears most accurate and setting the fixed offset as the difference between the estimates at this point.

We show quantitative results from the tracking in Figure 2. Selected frames from the tracking procedure compared with the equivalent estimate from our comparison method are shown in Figure 3.

	Mean Error (mm)	Std. Dev. Error (mm)
X axis - Our Method	1.51	1.48
X axis - Comparison	1.73	1.21
Y axis - Our Method	1.25	1.04
Y axis - Comparison	1.89	1.17
Z axis - Our Method	3.05	2.68
Z axis - Comparison	9.86	4.89

Table 1: The numerical results showing the mean and std. dev. of error in each axis.

3.2 Qualitative results

We also demonstrate the qualitative results of our method by performing tracking on several sequences from surgical environments where 3D tracking data is not available. This dataset was not captured with a stereo camera which prevents us from incorporating these constraints in our pose estimation. Selected frames from this validation are shown in Figure 4.

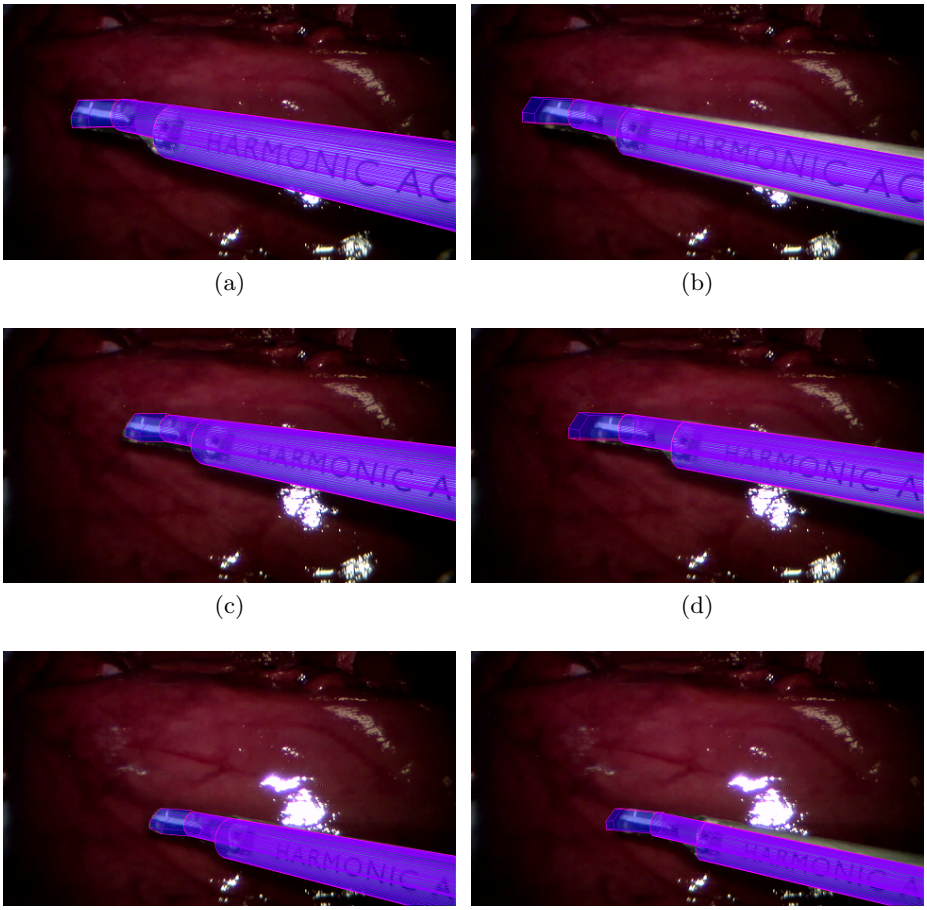
3.3 Failure modes

The most significant point of failure in our algorithm is dealing with a poor initialization, which is typically due to difficulties in correctly labelling the image pixels using the random forest. When this occurs, the model is placed too far from the ideal location for convergence to occur.

A secondary failure mode occurs due to our treatment of the instrument color model with a bag-of-pixels approach. This means that when the (often different colored) tip of the instrument is occluded behind tissue (e.g. due to cutting) the model can still fit to the image with a high degree of confidence as it doesn't care if the pixels it matches to the tip region of the contour actually match the true surface color at that point, only that they match the appearance model of the whole instrument surface. Potentially the appearance of the instrument model can be broken up into multiple classes [21] but as of yet this is not an area we have investigated.

4 Conclusion and Discussion

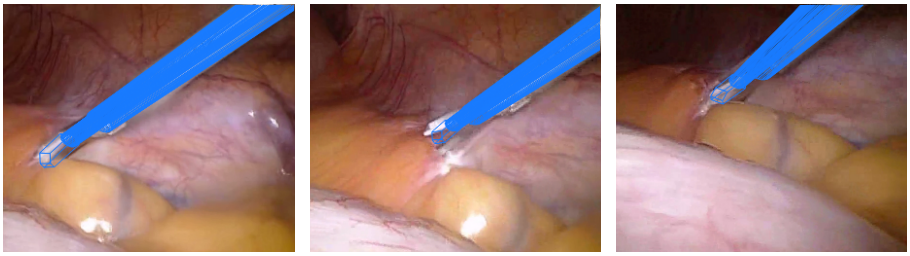
In this work, we have presented a novel framework for tracking rigid 3D objects using stereo 2D images. We combine a region based segmentation technique with point based pose estimation simultaneously addressing the weaknesses of both methods. Quantitative validation is performed on optically tracked endoscopic



(a) (b) (c) (d) (e) (f)

Fig. 3: The images show estimates of the instrument pose overlaid on the video. The left hand column of images show our technique which incorporates stereo, points and a Kalman filter compared with the right hand column showing the method of [14] which does not use these features.

images in a mock surgical environment. Figure 2 shows the estimated (x, y, z) position of the instrument tip compared with the method of [14]. Both methods provide good accuracy in x and y, although ours appears slightly more accurate and there is a significant accuracy improvement in the z direction, which is to be expected given the stereo constraints our method includes. The decrease in error over the duration of the sequence can be explained by the method gradually recovering from inaccuracies in the pose initialization. Table 1 shows the numerical performance improvements of our method. Visual comparison can be seen in sample frames in Figure 3 where both methods converge to an accurate solution but our method more accurately converges around the instrument tip and does not have the same errors in estimating the shaft rotation. The full video



(a) (b) (c)

Fig. 4: The frames show select examples from an *in vivo* dataset with the instrument model overlaid at the current pose estimate.

can be found online at `https://youtu.be/5VyRmvGBT8k`. Qualitative validation on *in vivo* data demonstrates that our method is feasible in real surgical environments. Sample frames showing the alignment accuracy are shown in Figure 4 which demonstrates the method's robustness to lighting and fast motion as well as the significant articulation in some frames.

Our method presents several areas where improvement is necessary. The most significant of which is the modelling of instrument articulation. Methods of disjoint optimization appear the most simple, where each articulated component is optimized separated however [21] and [22] have both presented methods of 3D pose tracking which handle the articulation as part of a single optimization. Additionally, further constraints need to be added to model the trocar insertion point which would help to improve the accuracy of our system.

Acknowledgements: The authors would like to thank CJMedical for supplying the Viking laparoscope used in the experiments. Danail Stoyanov would like to acknowledge the financial support of a Royal Academy of Engineering/EPSRC Fellowship. Max Allan would like to acknowledge the financial support of the Rabin Ezra foundation as well as the EPSRC funding for the DTP in Medical and Biomedical Imaging at UCL. John Kelly would like to acknowledge the UCL Biomedical Research Centre for their financial support.

References

1. H. Azimian, R. Patel, and M. Naish, "On constrained manipulation in robotics-assisted minimally invasive surgery," in *3rd IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics*, 2010, pp. 650–655.
2. E. P. Westebring van der Putten, R. H. M. Goossens, J. J. J., and D. J., "Haptics in minimally invasive surgery a review," *Minimally Invasive Therapy & Allied Technologies*, vol. 17, no. 1, pp. 3–16, 2008.
3. S. Speidel, G. Sudra, J. Senemaud, M. Drentschew, B. P. Müller-Stich, C. Gutt, and R. Dillmann, "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.
4. M. K. Chmarra, C. A. Grimbergen, and J. Dankelman, "Systems for tracking minimally invasive surgical instruments," *Minimally Invasive Therapy & Allied Technologies*, vol. 16, no. 6, pp. 328–340, 2007.

5. 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.
6. D. Stoyanov, "Surgical vision," *Annals of Biomedical Engineering*, vol. 40, no. 2, pp. 332–345, Feb. 2012.
7. D. R. Uecker, C. Lee, Y. F. Wang, and Y. Wang, "Automated instrument tracking in robotically assisted laparoscopic surgery," *Journal of image guided surgery*, vol. 1, no. 6, pp. 308–325, 1995, PMID: 9080352.
8. 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*, 2012, vol. 7511, pp. 568–575.
9. R. Richa, M. Balicki, E. Meisner, R. Sznitman, R. Taylor, and G. Hager, "Visual tracking of surgical tools for proximity detection in retinal surgery," in *IPCAI*, ser. IPCAI'11, 2011, pp. 55–66.
10. T. Zhao, W. Zhao, D. J. Halabe, B. D. Hoffman, and W. C. Nowlin, "Fiducial marker design and detection for locating surgical instrument in images," Patent US 068 395, 07 08, 2010.
11. 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, no. 11-12, pp. 1173 –1190, Nov. 2007.
12. A. Reiter, P. K. Allen, and T. Zhao, "Appearance learning for 3d tracking of robotic surgical tools," *The International Journal of Robotics Research*, 2013.
13. R. Austin, A. P. K, and Z. Tao, "Articulated surgical tool detection using virtually-rendered templates," in *Computer Assisted Radiology and Surgery*, 2012.
14. 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 biomedical engineering*, vol. 60, no. 4, pp. 1050–1058, 2013.
15. V. A. Prisacariu and I. D. Reid, "PWP3D: Real-Time segmentation and tracking of 3D objects," *Int. J. Computer Vision*, vol. 98, no. 3, pp. 335–354, Jan. 2012.
16. L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov, "Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery," *Medical Image Analysis*, vol. 17, no. 8, pp. 974 – 996, 2013.
17. D. Cremers, M. Rousson, and R. Deriche, "A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape," *Int. J. Comput. Vision*, vol. 72, no. 2, pp. 195–215, Apr. 2007.
18. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
19. S. Prince, *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012.
20. R. Tsai and R. Lenz, "A new technique for fully autonomous and efficient 3d robotics hand/eye calibration," *Robotics and Automation, IEEE Transactions on*, vol. 5, no. 3, pp. 345–358, 1989.
21. 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*, May 2009, pp. 3940 –3947.
22. V. A. Prisacariu and I. Reid, "Nonlinear shape manifolds as shape priors in level set segmentation and tracking," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 2185–2192.