Toward Detection and Localization of Instruments in Minimally Invasive Surgery

Article *in* IEEE transactions on bio-medical engineering · November 2012

DOI: 10.1109/TBME.2012.2229278 · Source: PubMed

Towards Detection and Localisation 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*—Methods for detecting and localising surgical instruments in laparoscopic images are an important element of advanced robotic and computer assisted interventions. Robotic joint encoders and sensors integrated or mounted on the instrument can provide information about the tool's position but this often has inaccuracy when transferred to the surgeon's point of view. Vision sensors are currently a promising approach for determining the position of instruments in the coordinate frame of the surgical camera. In this study, we propose a vision algorithm for localising the instrument's pose in 3D leaving only rotation in the axis of the tool's shaft as an ambiguity. We propose a probabilistic supervised classification method to detect pixels in laparoscopic images that belong to surgical tools. We then use the classifier output to initialise an energy minimisation algorithm for estimating the pose of a prior 3D model of the instrument within a level set framework. We show that the proposed method is robust against noise using simulated data and we perform quantitative validation of the algorithm compared to ground truth obtained using an optical tracker. Finally, we demonstrate the practical application of the technique on *in vivo* data from MIS with traditional laparoscopic and robotic instruments.**

I. Introduction

Reducing the access trauma of interventional healthcare through Minimally Invasive Surgery (MIS) is a paradigm of modern surgical practise [1]. By inserting specialised instruments and a laparoscope through small access ports, the surgeon can operate on the internal anatomy under direct video observation of the surgical site without making large incisions. The reduction in trauma has numerous benefits for the patient but impeding direct access to the surgical site complicates dexterous instrument-tissue interactions and makes procedures more difficult to perform. To overcome some of the challenges of MIS, robotic and computer assisted systems have been developed to enhance the control and navigation capabilities of the surgeon [2], [3], provide high resolution video feedback and eliminate hand tremor. However, with the increasing use of MIS in complex procedures new challenges and demands are emerging and for advanced computer assisted surgery it is important to integrate enhanced instrument constraints and image-guidance to protect critical structures and help the surgeon to locate anatomical targets. These systems are reliant upon real-time knowledge of the position of the instruments within the surgical field of view. This can be provided by robotic encoders and external tracking systems but these systems are often corrupted by whole system calibration errors meaning that detection and tracking of surgical instruments in MIS remains a challenge.

Video-based detection and tracking of surgical instruments in MIS has the advantage of being able to localise the tools within the surgeon's field of view without the need for additional hardware in the operating room. Probabilistic techniques using the natural difference in colour between the instruments and soft-tissues anatomy have been proposed for visual servoing in a robotic laparoscope holder [4] using naive Bayesian filtering. By augmenting the instruments with infrared or colour-based markers the problem can be simplified for laparoscopic [5], [6] and articulated robotic tools. However, such techniques require specialised manufacturing for *in vivo* use and the customised markers can often become obscured by occlusions and bleeding which then poses a significant challenge intra-operatively. Alternatively, [7] localised the instrument by using gradients filtered for consistency with the known insertion point of the trocar. Although the gradient approach can suffer from noise and shadows, using the trocar constraint has resulted in a successful 3D tracking system using the condensation algorithm [8]. Using the trocar constraint provides robustness, however, some practical considerations such as trocar movement with insufflation variation or patient movement may complicate the method in clinical use. More recently, techniques for handling the articulated tools of robotic surgery have been presented. A method of aligning an articulated model with a probability map using colour and simple texture features has been developed by [9] but the technique has high computational cost due to its use of the Nelder-Mead simplex method to optimise their cost function. Handling articulated instruments has also been addressed by [10] with a system that automatically recognises the instrument type and orientation from a learnt eigenspace representation. In a novel study, [11] used discriminative feature descriptors trained on specific regions of the robotic instrument to automatically detect points on instrument heads using a cascaded classifier approach. The spatial configuration of these features can then be used to estimate the pose of the instrument. The work has recently been refined further with promising results though the method largely relies on the da Vinci API after an initial pose estimate using stereo [12]. Another promising method with significant success is the use of brute-force template matching followed by pose refinement in [13]. An alternative approach from [14] combined detection and

tracking techniques within an entropy minimisation framework to determine the pose parameters of the surgical tools in retinal microsurgery. The results of the technique look promising for retinal surgery images but the method does not provide full 3D tracking of pose. In a another recent work in the field of retinal microsurgery [15] use gradient based 2D tracking to localise the instrument tip.

While literature on instrument detection and tracking is rapidly growing only a limited number of studies have addressed the difficult problem of instrument 3D pose estimation from monocular images [9]. In this study, we present a method of localising tool pose in 3D while achieving robustness to the large occlusions and lighting issues of MIS by using the strong prior of a known instrument shape. Our approach makes use of a region-based level set framework to incorporate this strong prior while maintaining a well behaved objective function that can be differentiated easily [16]. Because we rely on discriminative region statistics between the target object and the background we experiment with different image features including colour models and structural descriptors to perform tissue-instrument discrimination. Our approach can be used to recover instrument poses in 3D space up to 5DOF, as the axial motion around the shaft cannot currently be recovered. We report evaluation of our method on simulated data corrupted by varying degrees of noise and blurring to verify the numerical performance of our technique. Quantitative experiments were performed using instruments tracked by attaching optical markers for ground truth and results on images from MIS with both conventional and robotic instruments are provided for qualitative evaluation. The data used in our study and an executable of our method will be available online as part of our efforts to establish a benchmarking methodology to assist future developments in the field[1].

II. Appearance Learning with Random Forests

Random Forests (RF) [17] were the chosen method of classifying pixels in our images into either instruments or tissue. They provide an accurate, fast and potentially parallelisable classification method and offer an easy extension to multi-class data, a useful feature for classifying multiple distinct tool or tissue types [18]. The success of RFs has been due to their good generalisation ability which increases with the number of trees in the forest combined with the robustness to noise that randomness provides. This has led to their use in a variety of classification and regression computer vision problems such as keypoint recognition [19] and semantic image labelling [20].

A random forest is constructed as an ensemble of randomised decision trees, each of which consists of a set of weak learners that iteratively divide the classification of a sample $\mathbf{x}$ into a hierarchy of simpler problems. This is achieved by partitioning the sample space with decision boundaries and applying a different classifier in each region. Each applied classifier is either a decision node, which further partitions the search space and is represented as

$$h(\mathbf{x}, \boldsymbol{\theta_j}) \in \{0, 1\} \tag{1}$$

where $\boldsymbol{\theta_j}$ is a parameter vector which dictates the shape and position of the partitioning j^{th} hyperplane, or a leaf node which labels the sample as belonging to one of the desired classes

$$c^\star = \max_c p(c|\mathbf{x}) \tag{2}$$

where $c^\star$ is the labelling and $p(c|\mathbf{x})$ is the posterior probability of the class c given the sample.

A. Forest training

An RF is trained using the method of *bagging* which increases the generalisation of the resultant classifier by only training each member of the ensemble on a subset of the data by uniformly sampling with replacement. This is a method by which randomness is added to the forest because each single tree is trained on a slightly different set of data thus resulting in a different structure.

Each tree is grown incrementally by creating a new node then choosing a splitting parameter vector $\boldsymbol{\theta_j}$ which maximises an information gain type metric. This provides a further method of injecting randomness to the tree, as it is possible to randomly select a subset of all possible splitting parameters and maximise over these

$$\boldsymbol{\theta_j^\star} = \max_{\boldsymbol{\theta_j} \in \mathcal{T}_j} I_j \tag{3}$$

where $\mathcal{T}_j$ is a subset of the possible parameter vectors and I_j is the information gain type metric.

B. Features for Classification

For MIS images, due to the complexity of light reflectance in the scene, it is necessary to provide a thorough examination of different descriptors of the pixel data observed in each image. In the following section we explore several different features including different colour spaces and structural descriptors to determine which provides the most discrimination between the classes in our datasets. To ensure strong discriminative power over all regions of the instrument we split the class into 3 subclasses: the tip, dark components and light components. This ensures that when we choose features we select those which are distinctive in all regions of the tool, rather than selecting features which all classify strongly in the same region.

The colour spaces that we tested were RGB, HSV, CIE XYZ and opponent 2 (O2) and 3 (O3). Further details on the different color spaces are reported in [21], [22]. Figure 1 shows an illustration of some of the color spaces for a robotic surgery image.

In addition to color representations, we explored structural feature descriptors that are commonly used in the feature matching literature. We chose the gray-scale Scale Invariant Feature Transform (SIFT) [23] and colour-based SIFT descriptors [24] as well as Hessian of Gaussians (HoG) [25] implementations from the OpenCV library[2]. These are all

[1] http://www.cs.ucl.ac.uk/staff/m.allan/

[2] www.opencv.willowgarage.com

gradient based descriptors and work by concatenating vectors of gradient orientations around a particular pixel to provide a description of regions that can be invariant to contrast, intensity or small viewpoint changes [26]. The HoG descriptor groups gradients according to a specified window size, which allows the descriptor to detect on multiple scales. When selecting this parameter we chose to use the default size provided by the OpenCV library of 64×128 pixels.

C. Random Forest Implementation

We use a standard CPU implementation of Random Forests for classification of our video images. Our datasets consist of 6 sets each containing between 15 and 20 images from different minimally invasive procedures each with a different set of surgical tools. We found that by choosing the most visually contrasting frames from the procedure almost all of the feature variation within the images could be captured even by a small number of frames, this is due to the relative simplicity of our features. We split each dataset into a training set and a test set using two-fold cross validation as increasing the number of folds in the validation did not significantly affect the results. This technique enabled us to use a binary classification (tool or tissue) on each dataset rather than learning multiple instrument class labels, however, we hope to incorporate multiclass labelling in our classified such that the desired tool would be selected at run-time. We limited our forest size to 50 trees of no more than 10 levels to increase speed of both training and classification.

III. Instrument Pose Detection

A. Model Parametrisation

The following sections explain the method by which first the 2D pose then 3D pose of each instrument is computed. This involves a spatial parametrisation of the model which is achieved through the Euler angles (θ, ψ, ϕ) and a 3D translation vector $\mathbf{t}$ in the camera coordinate system.

At this stage of our investigation we have focused on localising an unarticulated instrument shaft which has full rotational symmetry and have not yet incorporated modelling of the tips. This is due to complications in both modelling the articulation of the tips as well as distinguishing the tip from the shaft during classification. This means that the axial rotational parameter ϕ is unrecoverable and as such our model has only 5DOF. It would be possible to capture this final degree of freedom by modelling a fully articulated instrument (where for example the clasper orientation could be used to determine ϕ) or a textured model approach.

B. Model Initialisation

Due to our prior knowledge of the shape of the surgical instrument, we can estimate its 2D pose using a method similar to [5]. We first isolate the largest connected regions C in a single binary image Ω (as seen in Figure 2b) with a flooding algorithm before performing shape analysis with the moment of inertia tensor of each region.

Each component of the moment of inertial tensor, I, is computed as:

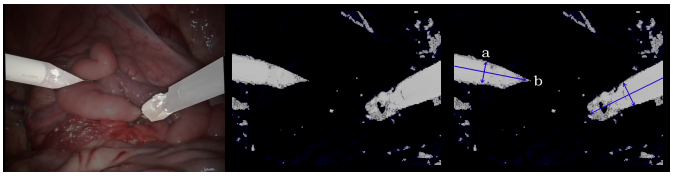
$$I_{ij} = \sum_{\mathbf{p}\in\Omega} [\mathbf{p} \in C](k^2\delta_{ij} - k_i k_j) \tag{4}$$

where $\mathbf{k}$ is the 2D vector from the cluster center $\mathbf{c}$ to the current pixel, $\mathbf{p}$, $k = |\mathbf{k}|$ and $i, j \in x, y$. δ_{ij} is the Kronecker delta.

The eigenvectors of this tensor give the principal axes of the shape as shown in Figure 2c as $\mathbf{a}$ and $\mathbf{b}$. Under the assumption that the tool can be approximated with solid cylinder, an estimate of the radius r and length l can be computed as $r = \sqrt{2I_2/m}$ and $l = \sqrt{12I_1/m - 3r^2}$ where m is the number of pixels in the shape and I_1 and I_2 are the largest and smallest eigenvalues, respectively. This equations can be derived from the well known moment of inertia equations for a cylinder, examples of which can be found in [27]. As the radius of the tool is fully observed in pixels in the image and known a priori in metric units (ignoring unlikely situations where the tool is parallel to and intersecting an image border) the newly computed radius r can be used to reverse engineer a rough starting estimate of the z-component of the vector $\mathbf{t}$ to the instrument center.

Finally, as we know the instruments must intersect the image boundaries, we filter the starting estimate of pose for irregularities. The position is checked for proximity to the image edge and ignored if one end of the tool is not within a minimum boundary of 20 pixels.

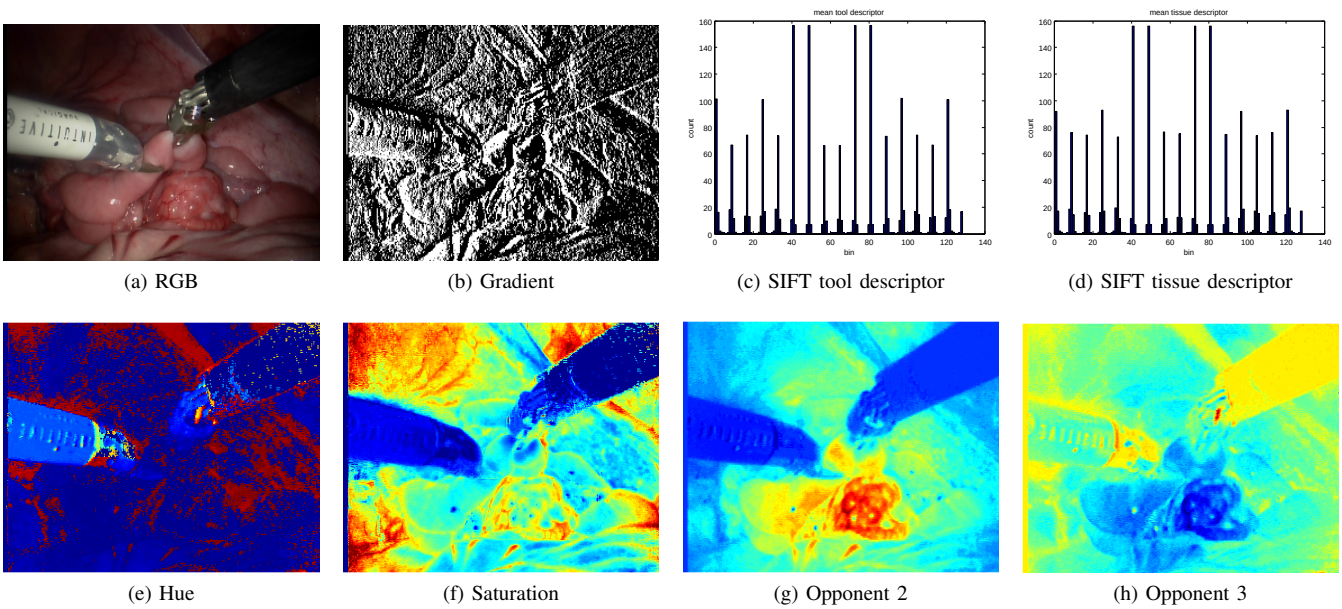
C. Refinement of pose



(a) Example & Mask (b) Forest classification (c) Principal axes

Fig. 2: (a) shows an example image from a surgical scene overlaid with a mask specifying which pixels belong to the tool and which to the background. (b) shows the output from the random forest, where each pixel intensity is the probability of the pixel belonging to the instrument class. (c) shows initialised principal axes of the tool from the segmented shape.

The model initialisation in 2D provides us with an estimate of the vector $\mathbf{t}$ to the tool center as well as the rotation angle θ between the x and y axes. To fully determine the remaining rotational degree of freedom ϕ as well as refine the starting estimates of θ and $\mathbf{t}$ we incorporate the segmentation and 3D pose recovery technique proposed by [28]. An image is segmented with a contour defined by the outer edge of a projected 3D instrument model, in our case a cylinder, at an estimated pose. By evolving this pose, which in turn changes the segmentation, an objective energy function which measures the pixel similarity between the interior and exterior regions and the learnt statistical models is minimised. The energy function is defined as:



(a) RGB (b) Gradient (c) SIFT tool descriptor (d) SIFT tissue descriptor

(e) Hue (f) Saturation (g) Opponent 2 (h) Opponent 3

Fig. 1: (a) shows the original RGB image. (b) shows an example of edge detection in the image, the noisy output demonstrates how challenging it is to use gradient information in the image to provide classification results. (c) and (d) show the average greyscale sift descriptor for both the tool and the tissue class. (e)-(h) show visualisations of the discriminative power of the chosen colourspaces. Parts of the instrument are clearly visible against the background, especially where the dark shaft is occluded by shadow in the RGB image, highlighting the advantage of this representation.

$$E(g(\mathbf{p})) = -\sum_{\mathbf{p}\in\Omega} \log\Big(H_e(g(\mathbf{p}))P_f + \big(1 - H_e(g(\mathbf{p}))\big)P_b\Big) \tag{5}$$

Here $P_i \;\; i \in \{f, b\}$ equates to a confidence value of a pixel $\mathbf{p}$ belonging to either of the two classes, tool or tissue and is a value returned by the method in section II-A as the proportion of trees which voted for each class. The level set function is represented as a signed distance function $g(\mathbf{p})$ representing the minimum distance between the pixel $\mathbf{p}$ and the contour defined by the zero level-set. This value is normalised with a smoothed Heaviside function, $H_e(g(\mathbf{p}))$.

This function has an intuitive interpretation in that the minimum value (because we are maximising a negative log) is reached when each pixel $\mathbf{p}$ with $P_f \approx 1$ is contained within the contour of the projected model (such that $H_e(g(\mathbf{p})) \approx 1$) and each pixel with $P_b \approx 1$ is outside the contour (such that $H_e(g(\mathbf{p})) \approx 0$). This situation occurs when the estimated model pose is best aligned with the class probability map.

In order to minimise the energy in line with transformations of the object pose, which we represent as an $\mathbb{SE}3$ transformation, the energy function is differentiated with respect to each of the 5 degrees of freedom λ_i:

$$\frac{\partial E}{\partial \lambda_i} = -\sum_{\mathbf{p}\in\Omega} \frac{P_f - P_b}{H_e(g(\mathbf{p}))P_f + (1 - H_e(g(\mathbf{p})))P_b} \frac{\partial H_e(g(\mathbf{p}))}{\partial \lambda_i} \tag{6}$$

$$\frac{\partial H_e(g(\mathbf{p}))}{\partial \lambda_i} = \delta_e(g)\left(\frac{\partial g(\mathbf{p})}{\partial x}\frac{\partial x}{\partial \lambda_i} + \frac{\partial g(\mathbf{p})}{\partial y}\frac{\partial y}{\partial \lambda_i}\right) \tag{7}$$

where $\delta_e(g) = \partial H_e/\partial g$ is a smoothed Dirac delta function and x, y are image coordinates in pixels. Although P_f and P_b are functions of the pixel coordinates $\mathbf{p}$, these derivatives are so small they can be treated as constants.

The contour is evolved using gradient descent, where the step size was heuristically set to 0.5° and 1.5mm, to account for the accurate initialisation we give our model.

IV. Experiments and Results

The proposed method was implemented in C++ on a machine running an Intel Core 2 Duo at 3.16 GHz without multithreading, GPU processing or significant code optimisation and hence our implementation is not currently real-time.

Classification of a single 720 × 288 image by the RF takes on average 3.47 seconds with pose localisation performance ranging from 1-40 seconds with the large variation in time being dominated by the initial starting solution. To validate the performance of the proposed method, we conducted experiments to evaluate different descriptor features for classifier training, we used synthetic data to test the pose estimation convergence and we show relative and qualitative results for *in vivo* videos. All datasets and their corresponding relative truth data are available online and we are packaging an executable of our method which will also be available to download.

A. Feature Selection for Classification

To evaluate the different colour and structural descriptors for the classification described in Section II we conducted experiments on the manually labelled data of 97 images. To select the most discriminative colour spaces, we used two different measures of dissimilarity: the variable importance

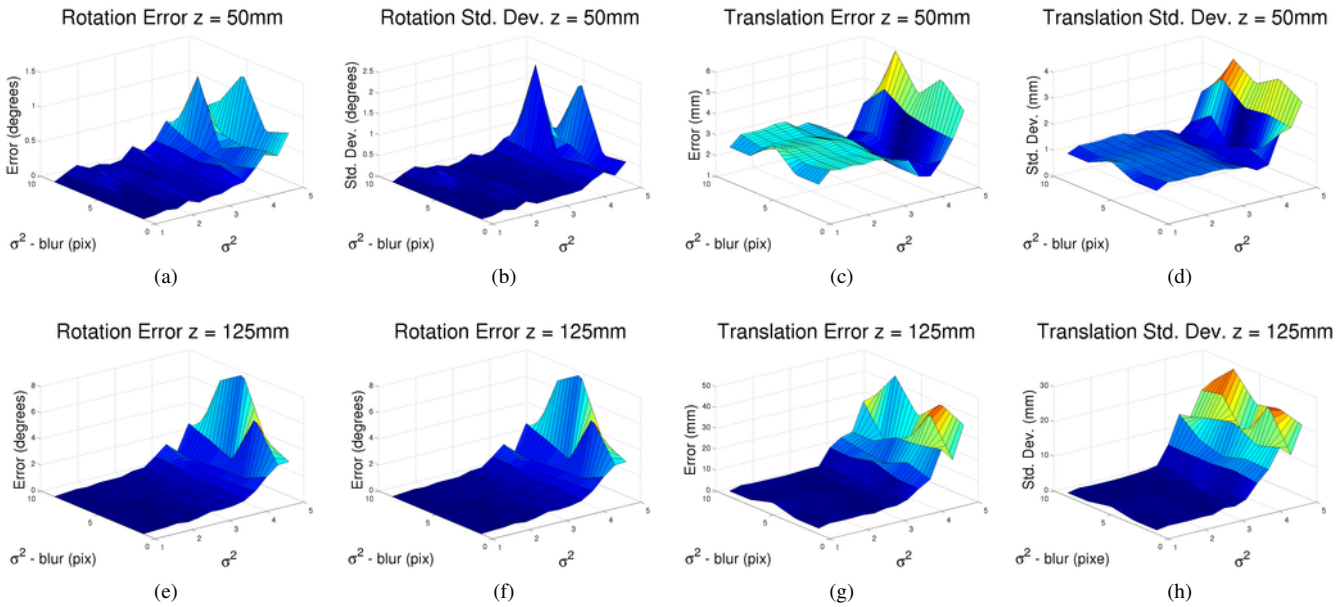


Fig. 3: Rotation and translation mean and standard deviation of error in the simulated data at depths of 50mm and 125mm with increasing image noise and blur. The light blue regions of the surface correspond to lower regions of error and the redder regions occurring at large levels of noise and blur correspond to higher pose errors. σ^2 blur refers to the variance of Gaussian kernel used for blurring and σ^2 refers to the variance of the Gaussian function used to create noise in the image.

and the Bhattacharyya Distance (BD). The variable importance [29] is computed by training a forest on all possible variables and assigning each a score based on the number of times the tree splits against that variable. The BD is a more general statistical technique which gives a measure of the dissimilarity between two probability densities $p_1(z)$ and $p_2(z)$ as $-\log B$ where B is the Bhattacharyya coefficient

$$B = \int \sqrt{p_1(z)p_2(z)}\,\mathrm{d}z \tag{8}$$

When computing the BD between classes we are required to use a one-against-all method to accommodate our multi-class approach to the instrument.

	R	G	B	H	S	V
VI	0.082	0.054	0.031	0.084	0.073	0.032
BD tips	0.196	0.318	0.224	0.498	0.480	0.205
BD light	0.397	0.556	0.472	0.670	0.707	0.391
BD dark	0.692	0.522	0.529	0.545	0.445	0.661
BD tissue	0.235	0.254	0.141	0.461	0.329	0.220

TABLE I: The variable importance (VI) and BD between the classes for RGB and HSV where large values for the VI and BD correspond to more distinct spaces.

	X	Y	Z	O2	O3
VI	0.064	0.061	0.053	**0.185**	0.110
BD tips	0.245	0.283	0.233	**0.587**	0.275
BD light	0.509	0.512	0.482	**0.758**	0.698
BD dark	0.650	0.578	0.533	**0.856**	0.663
BD tissue	0.329	0.247	0.156	**0.644**	0.275

TABLE II: The variable importance (VI) and BD between the classes for CIE XYZ and the Opponent 2 and 3 spaces. The largest values are for each class is highlighted in bold.

Tables I and II show that the most discriminative colour spaces appear to be the hue, saturation and the opponent 2 and 3 colour spaces as all were selected by both the BD and the variable importance across the instrument classes and the tissue. We used the most discriminative colour spaces as the basis for structural descriptor evaluation for classification. Here we evaluate the discriminative power using only the BD (shown in the previous section to have similar selective power to the variable importance) to reduce computational time in computing the difference measures. The texture features were computed at each pixel of a half resolution version of each image in order to reduce the computational complexity of the investigation but we also found that it improved performance accuracy.

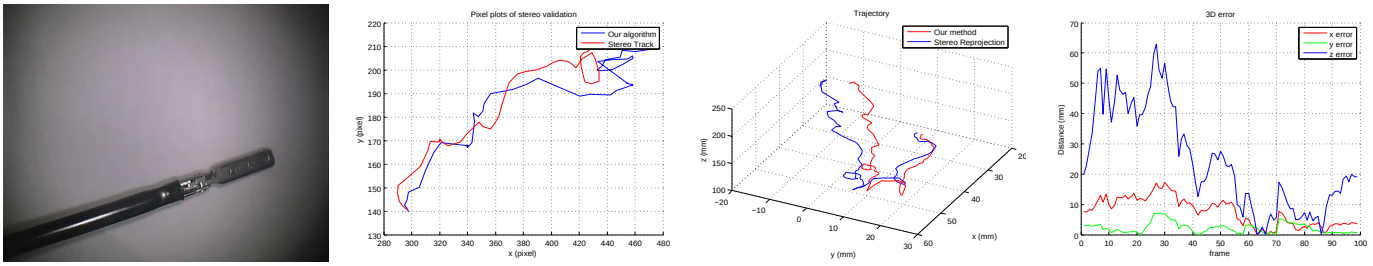
	RGB SIFT	Hue SIFT	Sat. SIFT	O2 SIFT	O3 SIFT
BD	**0.03**	0.02	**0.03**	0.02	0.02

TABLE III: The BD between the mean SIFT descriptor of each class in the examined colour spaces. All distances are very small, especially in comparison to the colour based distances. The largest values are highlighted in bold.

	HoG	Hue HoG	Sat. HoG	O2 HoG	O3 HoG
BD	**0.05**	0.03	0.03	0.03	0.02

TABLE IV: The BD between the mean HOG descriptor of each class of the examined colour spaces. As in the case of the SIFT descriptors, the distance is very small. The largest value is highlighted in bold.

As each method produces high dimensional vectors of gradient orientations around a given point, the descriptors are more challenging to compare than the colour based methods



(a) Example image (b) 2D trajectory - image (c) 3D trajectory - world (d) Error

Fig. 4: We track the instrument moving in front of a monochrome background and estimate the motion of the tip of the instrument using a colour-based template tracker. (b) shows plots of the pixel location of the tracked points. (c) shows the 2D trajectory from the colour template tracker in red and the trajectory estimated by our method in blue. (d) shows the Euclidean distance between the two points at each frame.

which allowed easy binning into histograms for comparison. To create a single descriptor for each class we average the descriptor in each region giving us 2 128-vectors, one for each class. To compare these descriptor vectors with the B.D. we generate histograms with each element of the vector as a separate bin. Although this is not a comprehensively quantitative comparison of the histograms of each class it acts as a proxy indicator of whether descriptors are discriminative features for robust classification. Table III and IV shows that the texture descriptors, in their standard form do not appear to be capable of effectively distinguishing tools and tissue, and due to their significant computation time, they were not included in the feature vector used to train our RF. It is important to note that as shown in [18] descriptors can be used to identify specific points on the articulated tool but our findings suggest that they are not suitable for whole tool identification. Further work is required to determine whether texture descriptors would be suitable to deliniating the instrument-tissue boundary rather than the tool body itself.

B. Experiments with simulated data

We generated simulated data to assess the accuracy of the pose fitting by projecting a cylinder with known 3D pose into an image before performing pose estimation with our approach. The validation results are shown in Figure 3 using the mean and standard deviation of the translation and rotation error between the estimated pose and the known pose. The simulation was conducted over a range of different noise values and distances between the model centre and the camera. As the classification task in synthetic simulations is trivial resulting in the unrealistic scenario of a perfect classification map, we distort the image with normally distributed noise and blurring provided by a Gaussian kernel which creates misalignment of the initial model. The projection was performed 15 times with the mean and standard deviation of error at a camera to model centre distance of 50mm and 125mm. These depths were chosen to approximate the minimum and maximum depths of the working volume in coventional laparoscopic and robotic surgery cases.

The low errors ($< 1^\circ$ and < 5mm) at all but the highest noise and blurring levels in Figure 3 illustrate that our approach is potentially numerically robust to noise. This is primarily due to the initialisation procedure being largely unaffected by additive noise which is distributed evenly around the instrument. However, once sufficient noise is present to misleadingly create artificial shaft shapes the error can rapidly rise with false initialisation. While simulated data does not fully capture the complexity and challenges of the surgical scene our results indicate that the numerical stability of our method is promising.

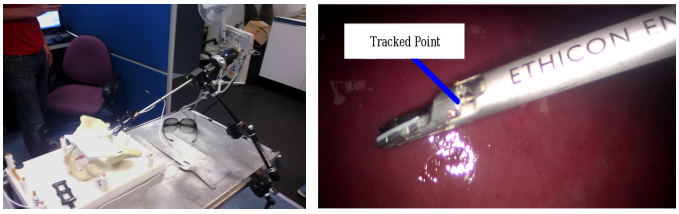
C. Laboratory experiments

To validate our method in a more realisitic and challenging environment than simulation we designed an experimental setup to test pose estimation against data acquired from an NDI Optotrak Certus[3] optical tracking system. We manufactured rigid bodies 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 5. 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 Optorak [45]. 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. We performed experiments by moving the tracked instrument in front of the camera covering a wide range of poses in front of an *ex vivo* lamb liver tissue sample. A capture workstation was used to synchronise the video and Optotrak readings giving us the pose for the instrument with respect to the camera coordinate system at each frame of video. 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 6. To compensate for calibration error, which results in a constant

[3]http://www.ndigital.com/lifesciences/certus-motioncapturesystem.php
[4]http://www0.cs.ucl.ac.uk/staff/Dan.Stoyanov/calib/
[5]http://www.vision.ee.ethz.ch/software/calibration_toolbox//calibration_toolbox.php

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. Incorporating a tracking framework to support our detection method would be an approach for reducing such errors. Furthermore by tracking and potentially locking onto a favourable depth may also enhance performance in the z direction.



(a) (b)

Fig. 5: (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.

D. Validation using in vivo datasets

We additionally perform pose validation on *in vivo* data using 97 hand segmented frames from 6 different surgical procedures described in section IV-A. Due to appearance differences between the instruments we treat each dataset separately training a unique classifier for each using half of the images as a training set and half as as testing set before swapping the sets. We compute a pose accuracy measure using the overlap between the projected instrument pose and the hand segmented frames as described by other authors [9], [13]. To obtain a data set that is representative of a full range of possible instrument poses we choose frames far apart in the sequence in which the tool pose was as unique as possible. The classification metrics we chose to use were true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) and calculate the average precision (P), recall (R) and probability of error (PE), equations for these values can be found in [9].

Results illustrating the performance of the RF classifier when compared to a Bayesian classifier (popular in surgical instrument tracking [30], [31]) are shown in table V where we show the measures comparing the pixel labelling from both classifiers compared to images with manually labelled pixels. The RF has better or at least almost equal accuracy on the data and has the additional advantage that is can easily be extended to work with multiple instrument classes or with different segments of the articulated surgical tool tip. The validation of our pose localisation method is shown in the results in Table VI. Using the hand segmented images and our estimate of the 3D pose we project the model of the instruments into each image. We then perform the binary accuracy measures using the pixels inside the projection's contour to represent positive pixels and pixels outside the contour to represent negative pixels.

	1	2	3	4	5	6
Bayes Precision	0.855	0.565	0.175	0.657	0.354	0.554
Bayes Recall	0.794	0.752	0.760	0.823	0.702	0.815
Bayes PE	0.018	0.019	0.130	0.044	0.040	0.050
RF Precision	0.840	0.530	0.317	0.716	0.390	0.500
RF Recall	0.840	0.753	0.697	0.834	0.700	0.840
RF PE	0.020	0.022	0.055	0.035	0.033	0.070

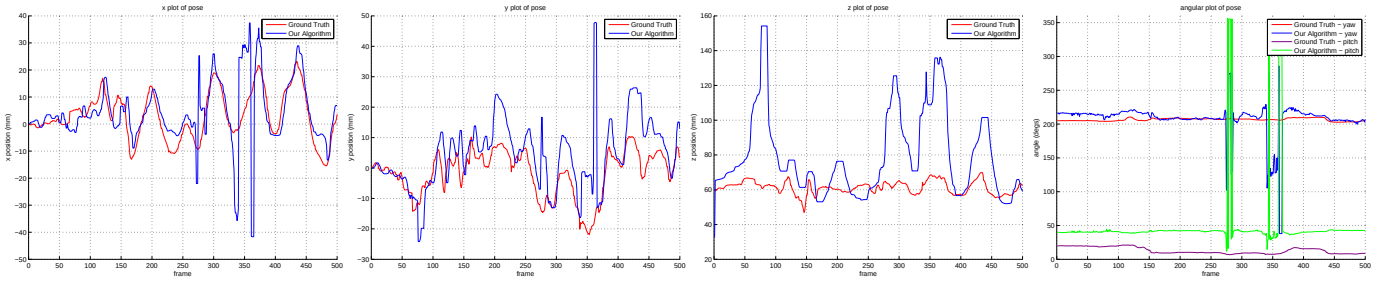
TABLE V: The binary accuracy measures when using a normal Bayesian classifier then an RF to label each pixel as belonging to instrument or tissue. The results are recorded across each of the 6 data sets. The results presented are obtained by averaging the scores across both folds of the 2-fold cross validation.

The precision and recall scores we achieve are promising as shown in Table V and VI and indicate that our technique has potential for good detection and subsequent pose estimation. Some datasets have better classification results due to the more distinctive tissue colouring where typically a deep red contrasts well with the limited colourfulness of the tool. Datasets containing fewer highly reflective surfaces or having moderate lighting conditions also perform better as we have found that highlights and shadows are typically misclassified as belonging to the instrument due to their similar achromatic appearance.

	1	2	3	4	5	6
Precision	0.910	0.801	0.633	0.785	0.705	0.597
Recall	0.694	0.660	0.503	0.783	0.703	0.681
PE	0.012	0.006	0.011	0.024	0.009	0.039

TABLE VI: The results for the binary classification methods when using the 3D pose estimation. The results are recorded across each of the 6 data sets.

While ground truth for 3D position and orientation is not available for *in vivo* data we performed preliminary validation of the 3D accuracy of the tool position estimation using stereo by tracking a region at the end of an instrument that is moving in front of a white background in front of the robotic stereo laparoscope. The cameras were calibrated with sub millimeter reprojection error (0.18 pixels) and we can reconstruct the tool tip point in 3D space using triangulation. We compared the position of this point to that of the current estimate of the instrument tip location estimated using our monocular technique. The quantitative results of the accuracy of the 3D position estimation are shown in Figure 4. The tool detection in the image performs well and the estimated position aligns with the stereo reconstruction and the majority of error as expected is in the z axis. The findings are promising and by



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

Fig. 6: 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.

incorporating stereo constraints we can improve the accuracy of our z axis estimation, however, our current monocular method is practical and does not require stereo. This is important as despite the inceasing use of da Vinci and other stereoscopic systems, monocular laparoscopic surgery is still the most widely adopted in clinical practice.

V. Discussion

In this study, we have proposed a method for detecting instruments in MIS images and estimating their pose in 3D camera space up to an ambiguity of 1 DOF in the instrument axis. Our detection method is based on classifying the laparoscopic image into instruments and background by using a trained classifier. After investigating different features to use for classification including colour models and structural descriptors, we found that the most effective classification results were obtained when using Hue, Saturation and Opponent 2 and 3 colour spaces. We fit an instrument model onto the classification using an energy minimisation algorithm and prior knowledge of the geometry of the instrument. In our current implementation the instrument model is a cylinder which accounts for the main tool shaft but this can be extended for the additional complexity of the tool tips and for articulated robotic instruments. The pose fitting algorithm uses the reprojection of the model to align to the classified image. With an initial solution provided by knowledge that the instruments enter the image from the outside (a constraint imposed by the MIS surgical setting) the convergence of the optimisation is efficient and fast. Our method is capable of dealing with large variations in noise shown in extensive testing with synthetic simulations. The method also performs well on images from surgical videos which we are making available to the community as a step towards establishing a benchmarking framework. Experiments with optical tracking as a validation metric indicate that our method is practical and although error in the z axis estimate can be improved on for a monocular method the algorithm performs well. Currently, our technique can be used to localize the 5 DOF pose of the surgical instruments from monocular views without rotation in the tool axis. To our knowledge this is one of the first methods capable of achieving this. We are actively working towards using additional model complexity at the tool tips to constrain this axial rotation. In our future work we will also investigate building stereoscopic constraints [32] which can increase robustness and also the 3D estimation accuracy. By incorporating tracking algorithms to apply temporal continuity to the tool localisation we expect to achieve improvements in performance which can be further optimized with parallelisation on GPU hardware aiming towards real-time performance.

Acknowledgements: The authors would like to thank Lena Maier-Hein for valuable comments about the work and 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] A. Darzi and S. Mackay, "Recent advances in minimal access surgery," *BMJ*, vol. 324, pp. 31–34, Jan. 2002.

[2] 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] D. Stoyanov, "Surgical vison," *Annals of Biomedical Engineering*, vol. 40(2), pp. 332–334, 2012.

[4] 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.

[5] O. Tonet, T. U. Ramesh, G. Megali, and P. Dario, "Tracking endoscopic instruments without localizer: image analysis-based approach," *Studies in Health Technology and Informatics*, vol. 119, pp. 544–549, 2006. PMID: 16404118.

[6] A. Krupa, J. Gangloff, C. Doignon, M. de Mathelin, G. Morel, J. Leroy, L. Soler, and J. Marescaux, "Autonomous 3-D positioning of surgical instruments in robotized laparoscopic surgery using visual servoing," *Robotics and Automation, IEEE Transactions on*, vol. 19, pp. 842 – 853, Oct. 2003.

[7] 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.

[8] R. Wolf, J. Duchateau, P. Cinquin, and S. Voros, "3D tracking of laparoscopic instruments using statistical and geometric modeling," *Medical Image Computing and Computer-Assisted Intervention: MICCAI ...*

International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 14, no. Pt 1, pp. 203–210, 2011. PMID: 22003618.

[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. Speidel, J. Benzeko, S. Krappe, G. Sudra, P. Azad, B. P. Muller-Stich, C. Gutt, and R. Dillmann, "Automatic classification of minimally invasive instruments based on endoscopic image sequences," in *In Proceedings of SPIE*, vol. 7261, 2009.

[11] A. Reiter, P. K. Allen, and T. Zhao, "Marker-less articulated surgical tool detection," in *Computer Assisted Radiology and Surgery*, 2012.

[12] A. Reiter, P. Allen, and T. Zhao, "Feature classification for tracking articulated surgical tools," 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. 592–600, Springer Berlin / Heidelberg, 2012.

[13] A. Reiter, P. K. Allen, and T. Zhao, "Articulated surgical tool detection using virtually-rendered templates," in *Computer Assisted Radiology and Surgery (CARS)*, 2012.

[14] R. Sznitman, R. Richa, R. H. Taylor, B. Jedynak, and G. D. Hager, "Unified detection and tracking of instruments during retinal microsurgery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, p. 1, 2012.

[15] 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.

[16] C. Bibby and I. Reid, "Robust Real-Time visual tracking using Pixel-Wise posteriors," in *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, (Berlin, Heidelberg), p. 831844, Springer-Verlag, 2008.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, p. 532, Oct. 2001.

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

[19] V. Lepetit and P. Fua, "Keypoint recognition using randomized trees," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, pp. 1465 –1479, Sept. 2006.

[20] P. Kontschieder, S. Bulo, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2190 –2197, Nov. 2011.

[21] 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.

[22] 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.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, pp. 91–110, Nov. 2004.

[24] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 1582 –1596, Sept. 2010.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, vol. 1, pp. 886 –893 vol. 1, June 2005.

[26] S. Prince, *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012.

[27] J. Taylor, *Classical Mechanics*. University Science Books, 2005.

[28] 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.

[29] 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.

[30] B. P. L. Lo, A. Darzi, and G.-Z. Yang, "Episode classification for the analysis of Tissue/Instrument interaction with multiple visual cues," *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003*, vol. 2878, pp. 230–237, 2003.

[31] S. Speidel, M. Delles, C. Gutt, and R. Dillmann, "Tracking of instruments in minimally invasive surgery for surgical skill analysis," in *Proceedings of the Third international conference on Medical Imaging and Augmented Reality*, Miar'06, (Berlin, Heidelberg), pp. 148–155, Springer-Verlag, 2006.

[32] D. Stoyanov, "Stereoscopic scene flow for robotic assisted minimally invasive surgery," in *Medical Image Computing and Computer Assisted Interventions (MICCAI12)*, pp. 479–486, 2012.