Towards real-time multiple surgical tool tracking

Maria R. Robu[a], Abdolrahim Kadkhodamohammadi[a], Imanol Luengo[a] and Danail Stoyanov[a,b]

[a]Digital Surgery, a Medtronic Company, 230 City Road, EC1V 2QY, London, UK;
[b]University College London, Gower Street, London, WC1E 6BT, London, UK

ARTICLE HISTORY
Compiled October 19, 2020

ABSTRACT
Surgical tool tracking is an essential building block for computer assisted interventions (CAI) and applications like video summarization, workflow analysis and surgical navigation. Vision-based instrument tracking in laparoscopic surgical data faces significant challenges such as fast instrument motion, multiple simultaneous instruments, and re-initialization due to out-of-view conditions or instrument occlusions. In this paper, we propose a real-time multiple object tracking framework for whole laparoscopic tools, which extends an existing single object tracker. We introduce a geometric object descriptor, which helps with overlapping bounding box disambiguation, fast motion and optimal assignment between existing trajectories and new hypotheses. We achieve 99.51% and 75.64% average accuracy on ex-vivo robotic data and in-vivo laparoscopic sequences respectively from the Endovis'15 Instrument Tracking Dataset. The proposed geometric descriptor increased the performance on laparoscopic data by 32%, significantly reducing identity switches, false negatives and false positives. Overall, the proposed pipeline can successfully recover trajectories over long-sequences and it runs in real-time at approximately 25-29 fps.

KEYWORDS
Multiple tool tracking; Surgical data science; Computer Assisted Interventions

1. Introduction

Computer assisted interventions (CAI) rely on the understanding and mapping of the surgical environment and surgical instrument tracking is a key CAI building block. The position and motion of the surgical instruments can enable skill analysis, phase detection, motion estimation, tool-tissue interaction and pave the way towards image guided interventions.

Recent approaches towards developing vision-based instrument tracking algorithms have focused on demonstrating feasibility in single object bounding box tracking [10, 6, 7]. A scale adaptive search strategy, as well as a probabilistic segmentation of background pixels were proposed as tools to increase tracking robustness during long sequences [10]. A recent comparison study of multiple vision-based single object trackers in minimally invasive surgery (MIS) data shows excellent performance on ex-vivo robotic sequences [6]. However, in-vivo laparoscopic videos lead to significant drops in performance. Due to their formulation, most single object tracking approaches

CONTACT Maria R. Robu Email: maria.robu@touchsurgery.com

cannot handle out of view conditions, multiple instruments and occlusions.

Alternative approaches focus on pose estimation of the surgical tools [21, 9] allowing for a more flexible representation than a bounding box. State of the art neural networks can be leveraged to localise instrument joints and model tool articulation [9]. Moreover, such tool parametrisation allows for multi-instrument disambiguation. Overall, current methods rely on additional information such as robot kinematics and 3D CAD tool models [21] which might hinder clinical translation, they cannot handle tool occlusion [21, 10, 9, 18] or recover temporal trajectories [21, 9] and real-time capabilities are prohibitive [10, 9].

Few tracking frameworks have been proposed to handle multiple tool trajectories over long sequences [18, 16]. For example, a weakly supervised neural network trained on frame-level presence labels has shown promising results localising tool tips [16]. They also present a quantitative evaluation of long-term trajectory information in cholecystectomy surgeries. However, due to the formulation based on presence labels, their approach cannot handle multiple similar tools in the same frame.

Most importantly, the majority of proposed methods focus on tracking a point centered on the tool tip or the bounding box enclosing it [10, 6, 16]. While efficient, such approaches will fail due to occlusion during tool-tissue or tool-tool interactions. However, full surgical instrument tracking is not straight-forward because tools tend to be rigid, elongated shapes that triangulate at the surgical site and bounding boxes are not effective labels for them, with the background anatomy covering the majority of the bounded area. MIS also involves multiple surgical tools entering and leaving the surgical view. Simply deploying multiple independent single object trackers for each tool would fail through drift accumulation, occlusions and poor scaling of computational complexity.

An interesting research direction would be to leverage the complementary strengths of the above methods. Single object trackers generally focus on robustness to scene appearance such as smoke, scale changes, illumination variation and fast motion of the tools, while being efficient. Alternately, pose estimation approaches can successfully disambiguate multiple overlapping tools and cope with occlusions due to their intuitive whole tool parametrisation.

In this paper, we formulate multiple-instrument tracking as a global optimization problem where the presence of more instruments only marginally increases computational time. We introduce a tracking manager, which maintains multiple simultaneous tool trajectories. We reformulate an existing tracker [20] to handle bounding boxes containing large areas of background, shared resources across trackers and increased efficiency. A novel geometric descriptor assists tool overlap disambiguation and fast motion handling. Experimentally, we provide an enhanced breakdown of tracking errors by adapting well-known multiple object tracking metrics CLEAR-MOT [4] to the surgical domain. Our proposed tracking pipeline was carefully designed to add functionality on top of key insights from previous promising techniques, while running in real-time (25-29 fps).

2. Methods

Our tracking pipeline fuses global (frame-level) to local (single object) scales to ensure real-time performance (see Fig 1). At a global scale, a binary segmentation model can be run at 1 fps in order to initialise new tracks. Since multiple promising solutions exist in the literature for surgical tool segmentation neural networks [17, 2], this paper

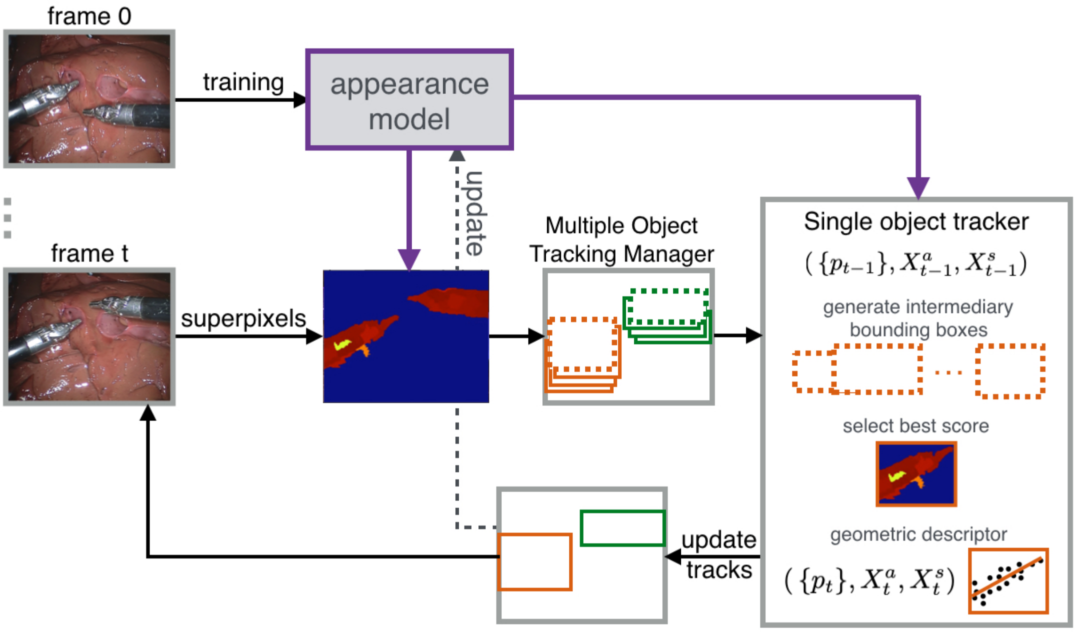
frame 0
training
appearance model
update
frame t
superpixels
Multiple Object Tracking Manager
Single object tracker
$(\{p_{t-1}\}, X^a_{t-1}, X^s_{t-1})$
generate intermediary bounding boxes
select best score
geometric descriptor
$(\{p_t\}, X^a_t, X^s_t)$
update tracks

Figure 1. Overview of the proposed tracking pipeline.

focuses instead on the tracking aspect. The segmentation maps can further provide labels for training a discriminative appearance model, alongside frame-level computation of superpixels and their associated feature vectors. The appearance model generates superpixel-based confidence maps, which localize tools for each frame. An online update scheme is employed to ensure the appearance model captures any changes in the environment and is able to learn from new frames.

At a local scale, single objects are tracked using a simple yet effective novel geometric descriptor. A tracker manager ensures long-term trajectories by leveraging the global confidence maps to generate candidate positions for each tracked object. The rest of this section introduces core insights from Yang et al. [20], followed by our proposed approach.

2.1. Single object tracker

Tracking is formulated using a Bayesian framework [20] where a motion model $p(X_t|X_{t-1})$ is combined with an observation model $p(Y_t|X_t)$:

$$p(X_t|Y_{1:t}) = \alpha p(Y_t|X_t) \int p(X_t|X_{t-1}) p(X_{t-1}|Y_{1:t-1}) dX_{t-1} \tag{1}$$

where α is a normalisation term. The aim is to estimate the object state X_t at frame t, given the previous observations $Y_{1:t}$. Given an observation Y_t, the likelihood of it belonging to the target or the background is approximated using a robust discriminative appearance model. Its role is to fuse low and mid-level structural information about the whole object, thus enabling recovery from drift and handling scale variation.

2.1.1. ***Appearance model***

A discriminative appearance model is trained based on the first n frames of the sequence. Given a bounding box at frame t, r superpixels are extracted $\{sp(t, r)\}$ [1], each represented by a feature vector f_t^r. After clustering the superpixel features, a score S_i^c is assigned to each cluster $clst(i)$, which indicates the likelihood of its members belonging to the target. Each resulting cluster $clst(i)$ is characterised by its center $clst_c(i)$, its radius $r_c(i)$ and its own members $\{f_t^r | f_t^r \in clst(i)\}$.

Yang et al. [20] considered all superpixels inside the active bounding box as the target area, which is a common assumption implying the region of interest/ bounding box contains very few background pixels [10]. When tracking whole tools, bounding boxes contain numerous background pixels, breaking the above assumption. Moreover, previous work [20] requires a user to manually segment a bounding box around the object in the first 4 frames of a video to extract positive and negative samples for the appearance model training. Such user intervention would not be feasible in the context of surgical tool tracking. We introduce a binary segmentation model as a solution to both these issues. Segmentation masks can be obtained automatically, which ultimately will be used to extract labels to inform the appearance model of the target location. As a result, a fully automatic framework tailored to tracking surgical tools can be achieved.

Finally, the trained appearance model is represented by the set of clusters and their associated cluster scores $\{(clst(i), S_i^c)\}$. With every new frame, a confidence map is estimated for an active bounding box by propagating cluster scores at the pixel level. High confidence values show high likelihood of target presence (see Fig. 1).

To adapt to changes in the scene, a sliding window update scheme of size H is used where a new instance is added every U frames as a circular buffer. An instance comprises of the bounding box and superpixels at frame t. The appearance model is updated every W frames by recomputing the clusters and associated scores $\{(clst(i), S_i^c)\}$ for the new collection of superpixels and labels [20]. In our experiments, we empirically selected $H = 5$, $U = 25$ and $W = 50$.

2.1.2. ***Efficient implementation***

We represent the motion model in Eq. 1 by generating a range of potential object locations with intermediary scales from the bounding box at state $t - 1$ and the matching candidate bounding box at state t. Each bounding box is scored by aggregating the corresponding confidence map values and the best one is chosen (Fig. 1).

Feature vectors f_t^r are extracted from each superpixel $sp(t, r)$, computed by concatenating the average RGB channels and average Cr, a and S from the YCrCb, Lab and HLS colour spaces respectively (6 bins). A variation of KMeans [13] is used to cluster the superpixel descriptors $\{f_t^r\}$ as we found KMeans (k=30) generalises well across different videos and helps fast confidence map estimation. For efficiency, superpixels and the confidence map are computed at the frame level and as a result, all trackers share the same appearance model and update strategy.

2.2. ***Internal object representation***

Similar appearance across multiple tools can make it difficult to disambiguate when their corresponding bounding boxes overlap. Instead, we propose to leverage additional geometrical information and temporal constraints.

The state X_t of a tracked object is formulated as $X_t = (\{p_t\}, X_t^a, X_t^s)$. Given a bounding box (encoded as width and height in X_t^s), a set of points $\{p_t\}$ are sampled from the bounding box using the confidence map as the sampling probability. Then, the principal axis of the surgical tool X_t^a is estimated from $\{p_t\}$ using principal component analysis. This assumes a single tool in the region, however, practically bounding boxes from multiple tools generally overlap. Also, changes in the scene due specularity can lead to false negatives in the confidence map. To handle these cases when updating the object tracker, prior information from state X_{t-1} is used to ensure the set of points p_t does not contain outliers.

Let $\{c_t\}$ be the initial set of candidate points at frame t. Two sequential pruning strategies are proposed. The first stage is based on the intuition that the set of points $\{p_t\}$ at state t will not differ significantly from state $t-1$.

$$\{p_t\} = \{c_t^i\} * \delta(d_N(\{p_{t-1}\}, \{c_t^i\})) * \delta(d(X_{t-1}^a, c_t^i)) \tag{2}$$

where $d_N(x, y)$ represents the distances from each element in X to its nearest neighbour in Y. Secondly, the axis of the tool at state t should be similar to X_{t-1}^a. The distances from points $\{c_t^i\}$ to the previous axis orientation X_{t-1}^a can be computed as:

$$d(X_{t-1}^a, c_t^i) = d((c_t^i - \overline{p_{t-1}}) - < (c_t^i - \overline{p_{t-1}}), X_{t-1}^a > X_{t-1}^a) \tag{3}$$

where $\overline{p_{t-1}}$ represents the centroid of the points at state $t-1$ and $< \cdot, \cdot >$ represents the dot product. The δ function then selects which candidate points are valid:

$$\delta(v^i) = \begin{cases} 1, v^i <= d_{max}(\mathbf{v}) \\ 0, otherwise \end{cases} \tag{4}$$

In both cases, $d_{max}(\mathbf{v})$ was empirically set to the $80th$ percentile of the sorted vector $\mathbf{v}$ for all experiments.

2.3. Multiple object tracking management

Tracked objects are initialized based on the connected components of the binary segmentation map with the geometric descriptor X_t and a unique trajectory ID. With each new frame, the MOT manager ensures candidate bounding boxes either match an existing trajectory or initialize new tracked objects. If there are no candidate bounding boxes, any existing tracked objects are marked as disappeared for 2 frames, after which they are deleted. Reliable data assignment between any candidate bounding boxes and existing tracks is critical.

Such matching can be formulated as a minimization of the total distance error between candidate objects C_t and existing tracks X_{t-1}. The Munkres algorithm [15] was used to obtain the optimal solution with polynomial runtime complexity. The intersection over union (IoU) is commonly used as a similarity metric between bounding boxes in order to perform the matching. However, this approach fails during fast motion or object overlap. We propose a distance error, which leverages the geometric object representation in order to improve the track assignment:

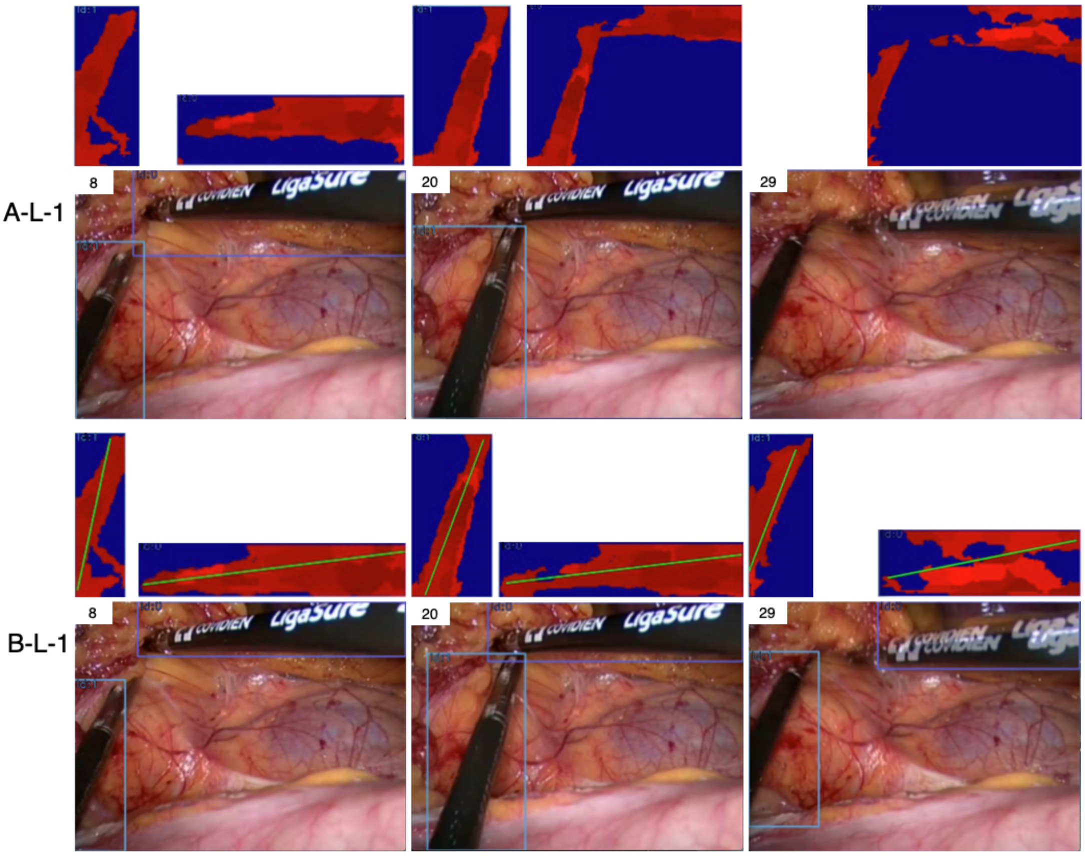

Figure 2. Successful tool disambiguation in sequence L-1. Comparison of tracking without (A) and with (B) the geometric object representation. Confidence map values corresponding to each tracked object are placed above each frame, where high values (red) represent the target and low values (blue) the background. The estimated axis X_t^a is overlaid with light green for B-L-1.

$$D = \beta(1 - IoU(X_{t-1}^s, C_t^s)) + (1-\beta)(1 - < X_{t-1}^a, C_t^a >) \tag{5}$$

Thus, the data assignment will favour bounding boxes with high overlap and similar axis orientation, which can help disambiguate between multiple tools. In practice, β was empirically set to 0.7 in all experiments.

Fig. 2 highlights the importance of both the proposed geometric descriptor (Sec. 2.2) and distance error D (Eq. 5) in cases where the bounding boxes overlap. In row A, the candidate bounding box encompasses both tools when they intersect. Without any instrument axis information, the right-hand tracker gets updated with the candidate bounding box based solely on the IoU overlap. The left-hand tracker is deactivated since there are no remaining available matches. On the other hand, row B shows that using the proposed geometric descriptor in the tracker update as well as in the data assignment can successfully disambiguate multiple tools interacting.

3. Experimental Setup

We provide a detailed analysis of the performance of the proposed tracker as well as qualitative results on robotic and laparoscopic data.

3.1. Metrics

The main error sources in tracking include: false positives (FP) - drifting away from the target, ID switches (IDSW) - switching to tracking a different tool or false negatives (FN) - complete failure to track. Generally, surgical tool tracking studies report a single localisation error which does not reflect these failure modes. While having high accuracy is desirable, tracking can provide other information about the environment, such as tool trajectories. We move towards the CLEAR-MOT [4] benchmarking approach used in the vision MOT literature.

Namely, the tracking accuracy (MOTA) aggregates the main error sources over the the number of objects at each frame det_t. The tracking precision (MOTP) illustrates how well the exact position of the bounding box is estimated.

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum det_t} \tag{6}$$

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t m_t} \tag{7}$$

where $d_{i,t}$ is the distance between tracked object i and its corresponding candidate at frame t and m_t is the number of matches found at frame t.

While surgical tool tracking was previously evaluated using $MOTA$ and $MOTP$ in [16], we provide a more detailed breakdown of the individual sources of error leading to a more in depth understanding of algorithm performance. The quality of the estimated trajectories can be summarized with the following metrics: mostly tracked (MT), partially tracked (PT) and mostly lost (ML). An object trajectory is considered *mostly tracked* if it was successfully tracked for more than 80% of its lifespan. If a track is recovered for less than 20% of its length, the target object is considered *mostly lost.* The remaining objects can be considered *partially tracked.* These metrics should be considered together with the ground truth number of trajectories (GT) in each sequence.

3.2. Data

We enhanced the existing Endovis'15 Instrument Sub-challenge [1] with annotations for whole tool bounding boxes and object trajectories for the sequences in the training dataset. Augmenting the robotic training data is straight-forward since segmentation masks are available at every frame. However, the trajectories in laparoscopic sequences were manually annotated since the ground truth segmentation masks were provided at 1 fps. Once an object leaves the frame and re-enters, a new ID is assigned - as proposed in [14].

[1] `https://endovissub-instrument.grand-challenge.org/`

Table 1. Quantitative evaluation without (A) and with (B) the geometric descriptor for robotic (R) and laparoscopic (L) sequences.

	MOTA ↑	MOTP ↑	IDS ↓	FP ↓	FN ↓	MT ↑	PT ↑	ML ↓	GT
A - R-1	97.47	72.34	3	9	44	2	0	0	2
A - R-2	99.29	86.26	0	4	4	1	0	0	1
A - R-3	100.00	75.91	0	0	0	1	0	0	1
A - R-4	98.93	78.29	0	6	6	1	0	0	1
A - Avg	*98.92*	*78.20*							
A - L-1	47.76	84.14	22	6	7	9	3	0	12
A - L-2	24.49	82.97	20	11	6	4	1	0	5
A - L-3	82.22	92.54	6	1	1	7	1	0	8
A - L-4	20.00	87.23	29	3	4	2	0	0	2
A - Avg	*43.62*	*86.72*							
B - R-1	100.00	70.36	0	0	0	2	0	0	2
B - R-2	99.29	82.23	0	4	4	1	0	0	1
B - R-3	100.00	78.16	0	0	0	1	0	0	1
B - R-4	98.75	72.46	0	7	7	1	0	0	1
B - Avg	*99.51*	*75.80*							
B - L-1	77.61	78.11	7	4	4	11	1	0	12
B - L-2	69.39	69.80	8	6	1	5	0	0	5
B - L-3	91.11	80.79	2	1	1	7	1	0	8
B - L-4	64.44	75.16	16	0	0	2	0	0	2
B - Avg	*75.64*	*75.97*							

Additionally, the provided segmentation masks are used as input in place of a binary segmentation model (Section 2) in order to isolate and evaluate the tracker's performance, uncoupled from any segmentation errors.

We use four 45 seconds sequences from the ex-vivo robotic data and four 45 seconds sequences of in-vivo laparoscopic videos. The robotic data contains a sequence with multiple tools and some situations where the bounding boxes overlap as a result of the tools approaching each other. The laparoscopic dataset illustrates extremely challenging situations such as multiple tools, tool overlap, out of view, blood, smoke, tool occlusions, presence of surgical objects (i.e. meshes, clips) and fast motion. Table 1 shows the results obtained with the proposed tracking pipeline with and without the geometric descriptor respectively. The arrows indicate if a larger (↑) or smaller (↓) value is better.

3.3. Results

The proposed tracking pipeline achieved 99.51% average accuracy on ex-vivo robotic data when using the geometric descriptor compared to 98.92% average accuracy without. Note that the biggest improvement can be observed for R-1, the only sequence containing multiple instruments with bounding box overlap.

The proposed MOT achieved 75.64% average accuracy on the in-vivo laparoscopic sequences, compared to 43.62% without the integration of the geometric descriptor. Note the high number of ground truth tracks for the laparoscopic videos, which indicates tools frequently going in and out of the field of view.

Overall, the use of the geometric descriptor significantly decreased the number of identity switches, false positives and false negatives. The proposed multiple object

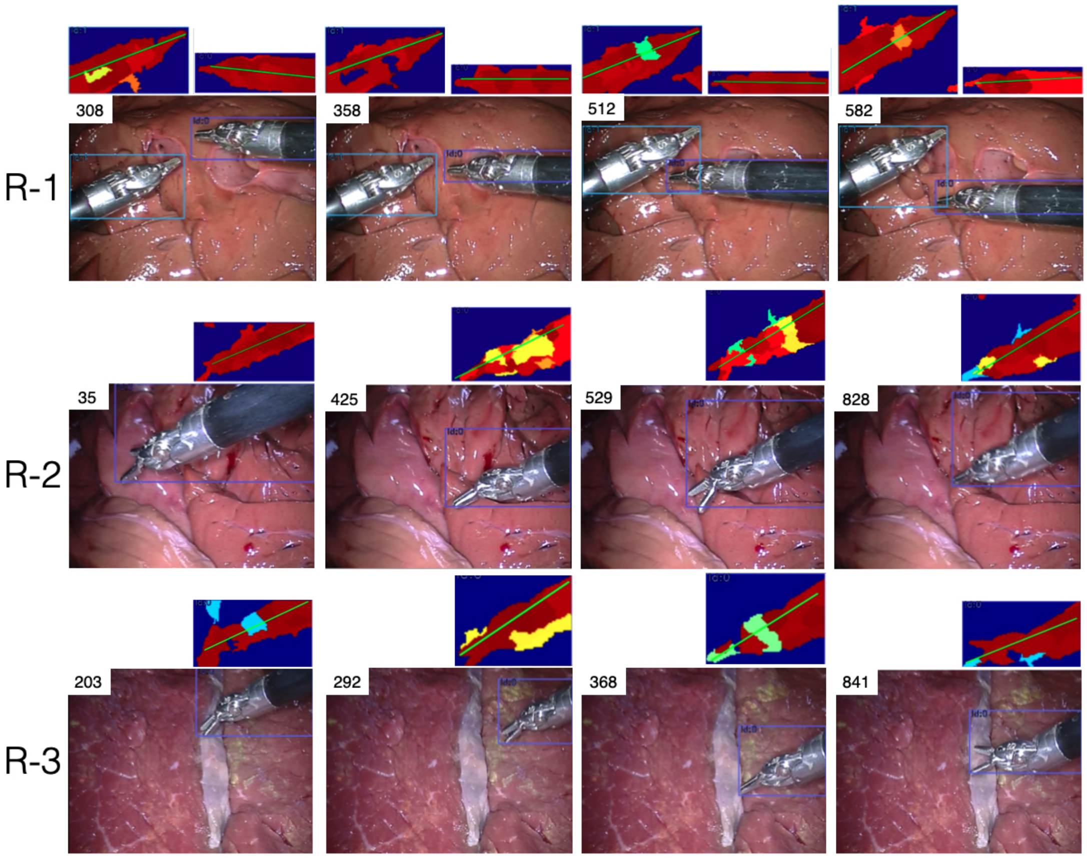

Figure 3. Proposed tracking on robotic sequences. The corresponding confidence maps for each tracker are overlayed with the estimated axis $\{X_t^a\}$ (light green) as well as the frame number in the upper-left corner.

tracker recovered more than 80% of the trajectory of all tools in the robotic data and of 25 out of 27 tracks on the laparoscopic data. Importantly, no tracks have been completely lost in either dataset.

Qualitative results are presented in Fig. 3 and 4. The tracker can handle multiple instruments (R-1), tool interaction with bounding box overlap (R-1), scale change (R-3, L-3), changes in appearance (R-2, L-1,3,4) and fast motion (L-1,3,4) over long sequences. Note the objects in Fig. 3 R-1 can be successfully tracked when their bounding boxes overlap due to the proposed estimation of the geometric descriptor. The last 2 columns of R-1 show that the distance and axis similarity based pruning of outliers correctly filters out any points selected on the neighbouring tool.

Direct comparison of errors with previous techniques is not possible due to differences in methodology (tracking tool tips vs. whole tool) and evaluation (different metrics) [7]. For example, a previous study achieves average MOTA and MOTP scores of 36.5% and 67.4% respectively while tracking tool tips in laparoscopic data [16]. However, most previous work focused on the feasibility and increased accuracy of single object tracking. Common metrics used consist of accuracy around a feature point on the tool tip or percentage of frames where the object is localized within a distance threshold [6, 10]. An additional fixed penalty can be added to the overall accuracy metric for each frame with tracking failure. While having high accuracy is desirable, when moving towards multiple object tracking, other information about the environment is needed - i.e. tool trajectories, a detailed breakdown of where and why tracking

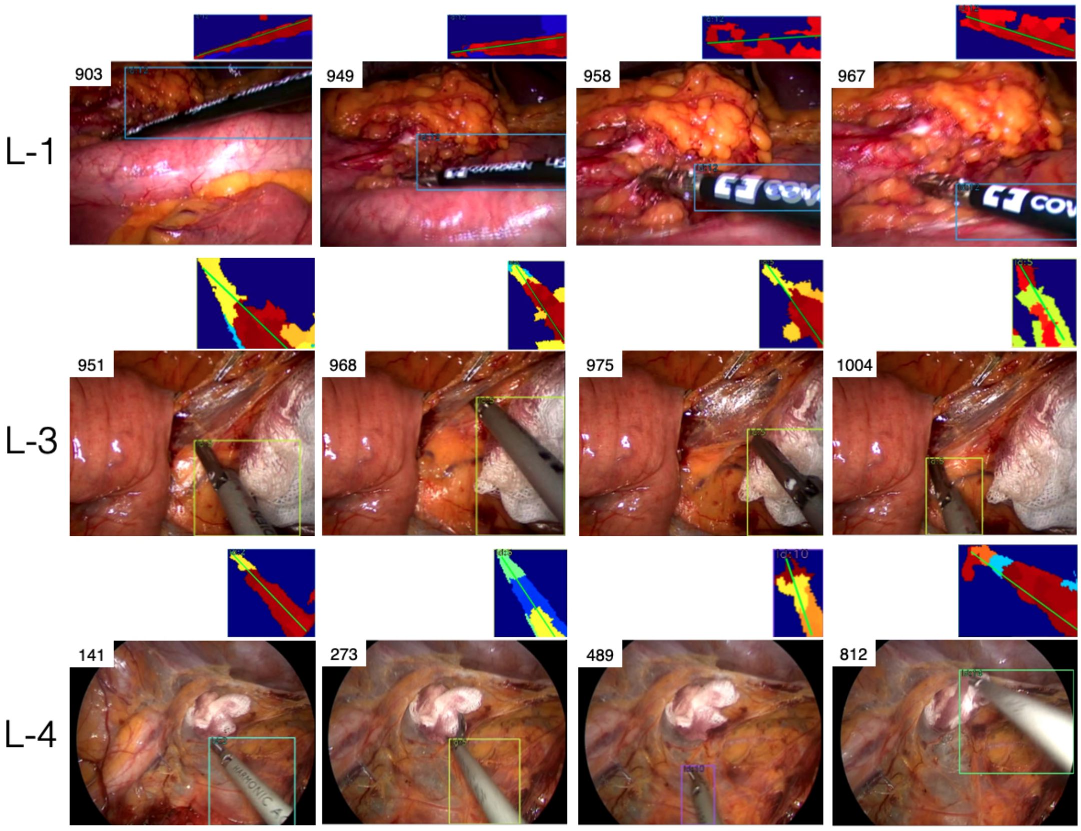

Figure 4. Proposed tracking on laparoscopic data (L-1,3,4). The corresponding confidence maps for each tracker are overlayed with the estimated axis $\{X_t^a\}$ (light green) as well as the frame number in the upper-left corner.

fails during long-term sequences. In surgery, particularly, there is a lack of maturity on datasets to benchmark tracking algorithms during long-term sequences with an increasing number of surgical tools. As future work, we plan to create such a dataset, similar to our previous work in segmentation: CaDIS [11].

4. Conclusion

We proposed a whole-object tracking pipeline for multiple laparoscopic instruments in MIS. Our results indicate that the proposed method is robust and can handle the challenges specific to MIS environments such as fast motion, tool interactions, multiple surgical tools and out of view conditions. We provided a tracking performance breakdown into sources of error, which could encourage the development of techniques tailored to specific clinical applications. For example, instrument pose accuracy would be critical to surgical navigation as opposed to workflow analysis applications, which would benefit more from good trajectories with a low number of identity switches. On the other hand, for applications with tool tissue interaction, having as few false negatives as possible might be more important.

Future work will include improving our appearance model to handle the initialization of new tracks at every frame. We will also look into making the bounding box estimation smoother over time to remove some of the jittering effect introduced by the superpixels.

Alternately, the tracking by detection paradigm consistently reaches state of the art performance in multiple vision benchmarking datasets [5, 19, 3]. Such MOT techniques generally rely on a real-time highly accurate detector being run at every frame, while a tracker manager ensures long-term trajectories with optimal data assignment. Promising results have been recently reported in surgical instrument segmentation [12, 8, 17, 7] which could greatly contribute towards an improved multiple object tracker.

5. Supplemental material

A video of the proposed tracking on robotic and laparoscopic video sequences is provided as supplemental material.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, *SLIC Superpixels Compared to State-of-the-Art Superpixel Methods*, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (2012), pp. 2274–2282.

[2] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.H. Su, N. Rieke, and I. Laina, *2017 Robotic Instrument Segmentation Challenge* (2017), pp. 1–14.

[3] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, *Tracking without bells and whistles* (2019).

[4] K. Bernardin and R. Stiefelhagen, *Evaluating multiple object tracking performance: The CLEAR MOT metrics*, Eurasip Journal on Image and Video Processing 2008 (2008).

[5] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, *Simple online and realtime tracking*, Proceedings - International Conference on Image Processing, ICIP 2016-August (2016), pp. 3464–3468.

[6] S. Bodenstedt, M. Allan, A. Agustinos, X. Du, L. Garcia-Peraza-Herrera, H. Kenngott, T. Kurmann, B. Müller-Stich, S. Ourselin, D. Pakhomov, R. Sznitman, M. Teichmann, M. Thoma, T. Vercauteren, S. Voros, M. Wagner, P. Wochner, L. Maier-Hein, D. Stoyanov, and S. Speidel, *Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery* (2018).

[7] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, *Vision-based and marker-less surgical tool detection and tracking: a review of the literature*, Medical Image Analysis 35 (2017), pp. 633–654.

[8] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, *Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers*, IEEE Robotics and Automation Letters 4 (2019), pp. 2714–2721.

[9] X. Du, T. Kurmann, P.L. Chang, M. Allan, S. Ourselin, R. Sznitman, J.D. Kelly, and D. Stoyanov, *Articulated multi-instrument 2-d pose estimation using fully convolutional networks*, IEEE Transactions on Medical Imaging 37 (2018), pp. 1276–1287.

[10] X. Du, M. Allan, S. Bodenstedt, L. Maier-hein, S. Speidel, A. Dore, and D. Stoyanov, *Patch-based adaptive weighting with segmentation and scale (PAWSS) for visual tracking in surgical video*, Medical Image Analysis 57 (2019), pp. 120–135.

[11] M. Grammatikopoulou, E. Flouty, A. Kadkhodamohammadi, G. Quellec, A. Chow, J. Nehme, I. Luengo, and D. Stoyanov, *CaDIS: Cataract Dataset for Image Segmentation* (2020), pp. 1–8, Available at `http://arxiv.org/abs/1906.11586`.

[12] Y. Jin, K. Cheng, Q. Dou, and P.a. Heng, *Incorporating Temporal Prior from Motion Flow for Instrument Segmentation in Minimally Invasive Surgery Video*, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 2019, pp. 440–448.

[13] S. Lloyd, *Least squares quantization in pcm*, IEEE Transactions on Information Theory 28 (1982), pp. 129–137.

[14] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, *MOT16: A Benchmark for Multi-Object Tracking* (2016), pp. 1–12.

[15] J. Munkres, *Algorithms for the assignment and transportation problems*, Journal of the society for industrial and applied mathematics 5 (1957), pp. 32–38.

[16] C.I. Nwoye, D. Mutter, J. Marescaux, and N. Padoy, *Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos*, International Journal of Computer Assisted Radiology and Surgery 14 (2019), pp. 1059–1067.

[17] T. Ross, A. Reinke, P.M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D.M. Filimon, P. Scholz, T.N. Tran, P. Bruno, P. Arbeláez, G.B. Bian, S. Bodenstedt, J.L. Bolmgren, L. Bravo-Sánchez, H.B. Chen, C. González, D. Guo, P. Halvorsen, P.A. Heng, E. Hosgor, Z.G. Hou, F. Isensee, D. Jha, T. Jiang, Y. Jin, K. Kirtac, S. Kletz, S. Leger, Z. Li, K.H. Maier-Hein, Z.L. Ni, M.A. Riegler, K. Schoeffmann, R. Shi, S. Speidel, M. Stenzel, I. Twick, G. Wang, J. Wang, L. Wang, L. Wang, Y. Zhang, Y.J. Zhou, L. Zhu, M. Wiesenfarth, A. Kopp-Schneider, B.P. Müller-Stich, and L. Maier-Hein, *Robust Medical Instrument Segmentation Challenge 2019* (2020).

[18] J. Ryu, J. Choi, and H.C. Kim, *Endoscopic Vision-Based Tracking of Multiple Surgical Instruments During Robot-Assisted Surgery*, Artificial Organs 37 (2013), pp. 107–112.

[19] N. Wojke, A. Bewley, and D. Paulus, *Simple online and realtime tracking with a deep association metric*, Proceedings - International Conference on Image Processing, ICIP 2017-September (2018), pp. 3645–3649.

[20] F. Yang, H. Lu, and M.H. Yang, *Robust superpixel tracking*, IEEE Transactions on Image Processing 23 (2014), pp. 1639–1651.

[21] M. Ye, L. Zhang, S. Giannarou, and G.Z. Yang, *Real-Time 3D Tracking of Articulated Tools for Robotic Surgery*, Lecture Notes in Computer Science, Vol. 9900, Springer International Publishing, Cham, 2016, pp. 386–394.