Surgical Tool Tracking by On-line Selection of Structural Correlation Filters

Daniel Wesierski[1,2] and Anna Jezierska[2,1]

[1] Gdansk University of Technology, Faculty of Electronics, Telecommunications, and Informatics
ul. Narutowicza 11/12, 80-233 Gdansk, Poland
[2]Systems Research Institute of Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland

***Abstract*—In visual tracking of surgical instruments, correlation filtering finds the best candidate with maximal correlation peak. However, most trackers only consider capturing target appearance but not target structure. In this paper we propose surgical instrument tracking approach that integrates prior knowledge related to rotation of both shaft and tool tips. To this end, we employ rigid parts mixtures model of an instrument. The rigidly composed parts encode diverse, pose-specific appearance mixtures of the tool. Tracking search space is confined to the neighbourhood of tool position, scale, and rotation with respect to previous best estimate such that the rotation constraint translates into querying subset of templates. Qualitative and quantitative evaluation on challenging benchmarks demonstrate state-of-the-art results.**

I. Introduction

Visual tracking of surgical instruments is an active research topic. It is a promising technology to be applied for virtual measurements and overlays [1], maintaining the operated tissue areas in view [2], guidance and recognition of risk situations [3], surgical skills assessment [4] and better understanding of surgical workflow [5]. Many methods have been proposed to deal with this challenging problem, whereas most of them concentrate on appearance term representing the whole tool. For review, we refer to [6].

The correlation filters based approaches aim at representing appearance of the target object using the learned templates. They are becoming popular due to low computational cost and robustness. Additionally, to improve tracking performance, recent approaches focus on exploiting structural information, e.g. relation among target parts. The structural information is usually used to improve the robustness in handling object deformation and partial occlusion.

For instance in [7],[8] the authors promote an object structure by introducing smooth constraint of confidence maps. In [9] the authors proposed to account for object transformation, translation change and scale variation by performing scale-spatial correlation jointly using a novel block-circulant structure for the object template and modeling target rotation in polar coordinate system. In [10] a structural correlation filter model is proposed, promoting a local change of scale, rotation and translation of all parts to be similar, while accounting for outliers part with different motion behaviour by introducing sparsity assumption. In [11] the authors proposed deformable part based correlation filter tracking. In this collaborative framework, the structure is represented by decomposing an object into several local parts, while deformation model is encoded by a global filter. In [12] the authors proposed not only to consider pairwise geometric relations between local parts, but to exploit high-order relations using the geometric hypergraph. Whereas in [13] the authors proposed to combine correlation filter framework with spatial-temporal angle matrix to account for object global rotation and deformation in visual tracking.

All above mentioned methods are designed for general tracking purpose. However, in some scenarios, the structure of the object to be tracked is known in advance, and it is possible to incorporate the specific prior knowledge when designing the tracker. Well studied examples of known in advance structure is a face recognition problem [14]. In this paper, we investigate a surgical tool. We make a use of its particular structure by decoding change of its pose between two subsequent frames into three relative rotation. To this end, we employ rigidly structured model of instrument parts proposed in [15] presented within adaptive tracking-by-detection framework [16]. The appearance is modeled within template-space representing HOG features. We introduce an injective function mapping each template related to shaft into two values coding rotation and scale of target object and each template related to end-effector into vector of four values coding relative rotations of tips and its scale. Using our model, a tracking search space is reduced to a local neighborhood of a tool position, scale, and rotation with respect to current state. A rotation constraint translates into the model in limited number of templates to be queried. We sample a resulting search-space with a pixel-wise dense patch sampling technique [17]. At each time step, both shaft and end-effector position, scale and orientation are updated. We propose a unified objective function to integrate these two sparse representation problems together. The function combines spatio-temporal constraints related to shaft and end-effector [18], [19], [20] with confidence maps of co-appearance of different part together. The resulting optimization problem can be well solved by efficient dynamic programming.

The remaining of the paper is organized as follows: In Section II, we present the general form of the addressed

ISBN 978-0-9928626-7-1

problem and introduce the notation used in this work. We then present the proposed method in Section III. Simulations are performed in Section IV, showing the good performance of the proposed approach. Finally, Section V concludes the paper.

II. Proposed problem formulation

A. Structural Maximum Margin Correlation Filters

At the frame t, given the target template set (dictionary) $D_{\mathcal{S}} = \left[d_{\mathcal{S}}^1, d_{\mathcal{S}}^2, \ldots, d_{\mathcal{S}}^{n_{\mathcal{S}}}\right]$ and $D_{\mathcal{E}} = \left[d_{\mathcal{E}}^1, d_{\mathcal{E}}^2, \ldots, d_{\mathcal{E}}^{n_{\mathcal{E}}}\right]$ related to Shaft parts, and End-effector poses, respectively, let $z^t = \{z^{1,t}, z^{2,t}, \ldots, z^{o,t}\}$ denote the observed, corresponding candidate target patch in target template space (e.g. HOG). In a correlation filters framework, a filtering/classification is based on searching a maximal correlation peak between observed data $z^{i,t}$ and a template $d_{\mathcal{S}}^i$. In the case of dictionary of templates the problem is to find a maximal correlation peak over all considered $z^{i,t}$ and all templates in dictionary $D_{\mathcal{S}}$ and $D_{\mathcal{E}}$.

Let $E^t = \{E^{1,t}, E^{2,t}, \ldots, E^{O,t}\}$ denote a set of blocks $E^{i,t}$ combining $N+1$ observed patches from z^t such that the first N patches altogether represent the complete structure of the shaft and the last one is representing end-effector. For all $E^{i,t} = \left\{\left(z^{1,t}\right)_i, \ldots, \left(z^{N+1,t}\right)_i\right\} \in E^t$ additional constraints are imposed to enforce all shaft-related parts within one block to be encoded with the same single template from dictionary $D_{\mathcal{S}}$. Note that the combination of the templates in the part related template-space to be the same was enforced in [21], where the problem was defined within sparsity tracking framework. In the following we limit our investigation to correlation filter framework. In the proposed approach a structure of a tool is controlled by: (a) linear coincidence of position of $N+1$ patches formatting each block $E^{i,t}$ and (b) representation of shaft by the same template. In addition we introduce in our model the learned biases matrix $B \in \mathbb{R}^{n_{\mathcal{S}} \times n_{\mathcal{E}}}$ favouring solutions combining reliable co-appearance of end-effector and shaft poses. For joint learning of dictionaries and biases we adopt a maximum margin correlation filters combining good localization properties of correlation filters with the very good generalization abilities of support vector machines [22]. For a detailed description of learning procedure we refer to [15].

With the selective assumption, the local blocks of patches within the considered set of blocks E^t can be represented through elements of the dictionaries by solving the following multi-criteria optimization problem:

$$\begin{aligned} \underset{x^t, y^t}{\operatorname{argmax}} \sum_{\substack{\left(z^{j,t}\right)_i \in E^{i,t}, \\ j=1,\ldots,N}} & \left(x^t\right)^\top D_{\mathcal{S}}^\top \left(z^{j,t}\right)_i + \\ & \left(y^t\right)^\top D_{\mathcal{E}}^\top \left(z^{N+1,t}\right)_i + \left(x^t\right)^\top B^\top y^t \\ \text{s.t} \quad & |x^t|_0 = 1, |y^t|_0 = 1 \end{aligned} \tag{1}$$

where $x^{i,t} \in \mathbb{R}^{n_{\mathcal{S}}}$ and $y^{i,t} \in \mathbb{R}^{n_{\mathcal{E}}}$ are selecting vectors with one non-zero coefficient equal to 1 related to a pose of single shaft part and end-effector part, respectively.

Note that in the above defined problem one variable x^t is assign to all parts related to shaft in given $E^{i,t}$, and hence the shaft is enforced to be represented by one template. One of the drawbacks of the problem formulation proposed in (1) is that it does not guarantee any temporary homogeneity of the resulting tracker. To alleviate this shortcoming, we propose to impose additionally dynamic constraints.

B. Geometrical model for tool dynamics

Templates $d_{\mathcal{S}}$ and $d_{\mathcal{E}}$ can be decried by its scale $a \in A$, where $A \in (0,1)^Q$ is a vector of considered scales. Additionally, let characterize a template $d_{\mathcal{S}}$ through an angel $r_1 \in (0, 2\Pi)$ in polar coordinate system with the origin fixed in a middle of shaft centreline and axis parallel to template box axis. Let introduce function $\theta_{\mathcal{S}}$ mapping $d_{\mathcal{S}}$ to its scale and above defined angle. Similarly let characterize a template $d_{\mathcal{E}}$ through a vector of two angles $[r_2, r_3] \in (0, 2\Pi)^2$ in polar coordinate system with the origin fixed on crossing point of centrelines related to shaft, left and right tips of microsurgical instrument (scissors) while axis parallel to template box axis. Let function $\theta_{\mathcal{E}}$ be a mapping $d_{\mathcal{E}}$ to its scale and a pair of above defined angles.

The dictionary $D_{\mathcal{S}}$ can be characterized by rotation/scale image $I_{\mathcal{S}} \in [(0, 2\Pi), A]^{n_{\mathcal{S}}}$, defined as: for every $d_{\mathcal{S}}^i \in D_{\mathcal{S}}, I_{\mathcal{S}}(i) = [r_1, a_1]_i \Leftrightarrow \theta_{\mathcal{S}}(d_{\mathcal{S}}^i) = [r_1, a_1]_i$. Similarly, the dictionary $D_{\mathcal{E}}$ can be characterized by rotation image $I_{\mathcal{E}} \in [0, 2\Pi]^{n_{\mathcal{E}} \times 2}$, defined as: for every $d_{\mathcal{S}}^i \in D_{\mathcal{S}}, I_{\mathcal{E}}(i) = [r_2, r_3, a_2]_i \Leftrightarrow \theta_{\mathcal{E}}(d_{\mathcal{E}}^i) = [r_2, r_3, a_2]_i$. We assume that the following properties hold:

- *The semantic structure of dictionaries* - The mapping from $D_{\mathcal{S}}$ to $I_{\mathcal{S}}$ and from $D_{\mathcal{E}}$ to $I_{\mathcal{E}}$ are injective, i.e. at given scale there is only one appearance template assign to given rotation of shaft and only one appearance template assign to given pair of rotations of end-effector.
- *The consistent motion property* - The displacements of target object's part is close within some distance measure in a space of: position of rotation center of the tool motion $p \in \mathbb{R}^2$ (for 2D case), considered rotations of shaft r_1, rotations of tips scissors r_2, r_3, and scales of shaft $a_1 \in A$ and end-effector $a_2 \in A$.

The relation between motion parameters and variables x, y can be expressed by function $\varphi : (\mathbb{R}^{n_{\mathcal{S}}}, \mathbb{R}^{n_{\mathcal{E}}}) \mapsto \left(\mathbb{R}^2, (0, 2\Pi), A, (0, 2\Pi)^2, A\right)$ defined as:

$$\varphi(x, y) = \begin{bmatrix} p \\ \theta_{\mathcal{S}}\left(D_{\mathcal{S}} x\right) \\ \theta_{\mathcal{E}}\left(D_{\mathcal{E}} y\right). \end{bmatrix}, \tag{2}$$

A dynamic state of surgical tool defined by its position and rotation was propoused in tracking approachby Zhou and Payande [23]. However in this work the authors considered much simpler and known to be less robust appearance model grounded only on edge extractions. Additionally, the authors did not use an information about structure of the tool, and hence only a global orientation parameter was used. In some other works the orientation of surgical tool was inexplicitly

ISBN 978-0-9928626-7-1 © EURASIP 2017

taken into account, for instance by examing variuous orientations of evaluation window [24].

C. Tracking framework

Our visual tracking problem is formulated within the Bayesian inference framework [7], [20] with spatio-temporal constraints [18], [19]. Similar to [23], we use the affine motion model with six parameters to describe the object's state $u^t = [p, r_1, a_1, r_2, r_3, a_2]$. In the following, we define our reconstruction problem in term of variables x^t, y^t and then we recover signal u^t form x^t, y^t using function φ defined in (2). In the context of inverse problems, the target state variable u^t can be recovered by minimizing a penalized criterion:

$$\underset{x^t, y^t}{\operatorname{argmin}} \quad \Phi(x^t, y^t) + \Psi(x^t, y^t) \tag{3}$$

$$u^t = \varphi(x^t, y^t) \tag{4}$$

where Φ is the so-called data fidelity term and Ψ is a regularization function incorporating a priori information, so as to guarantee the stability of the solution w.r.t. the observation noise. In the Bayesian framework, this allows us to compute the maximum a posteriori (MAP) estimate [25] of the original signal. In the context of our problem, the data fidelity term is given by observation model, i.e.

$$\Phi(x^t, y^t) = - \sum_{\substack{\left(z^{j,t}\right)_i \in E^{i,t}, \\ j=1,\ldots,N}} \left(x^t\right)^\top D_{\mathcal{S}}^\top \left(z^{j,t}\right)_i - \left(y^t\right)^\top D_{\mathcal{E}}^\top \left(z^{N+1,t}\right)_i \tag{5}$$

while regularization term incorporates dynamic and structural constraints, i.e. $\Psi(x^t, y^t) = \Psi_S(x^t, y^t) + \Psi_D(x^t, y^t)$, where

$$\Psi_S(x^t, y^t) = -(x^t)^\top B^\top y^t \tag{6}$$

$$\Psi_D(x^t, y^t) = \Omega\left(\varphi(x^{t-1}, y^{t-1}) - \varphi(x^t, y^t)\right) \tag{7}$$

and Ω denotes some distance measure.

III. Proposed algorithm

One of the drawbacks of the general problem formulation proposed in (3) is the computational complexity of a related solver. To alleviate this shortcoming, we propose to concentrate on an interesting special case, where Ω is defined as an indicator function $\iota_{\mathcal{C}}$, where

$$\iota_{\mathcal{C}}(u) = \begin{cases} 0 & \text{if } u \in \mathcal{C} \\ +\infty & \text{otherwise} \end{cases} \tag{8}$$

and $\mathcal{C}$ is a subset of $\left(\mathbb{R}^2, (-2\Pi, 2\Pi), \mathbb{R}, (-2\Pi, 2\Pi)^2\right)$. Such choice of distance measure is equivalent with an assumptions that in some neighbourhood defined by $\mathcal{C}$ around u^t all the solutions are equally probably (from a point of view of dynamic constraints) and outside this neighbourhood all the solutions are not possible. Hence, the constraints on $\left(p^{t-1} - p^t\right)$ translates into tracking search space to be reduced to some region "bounding box" around p^{t-1}, i.e. set of blocks E^t include only $E^{t,i}$ combining $N+1$ observed patches $z^{t,i}$ centred in a bounding box defined by $\left(p^{t-1} - p^t\right)$ and $\mathcal{C}$. Similarly, the constrained imposed on $[r_1, a_1, r_2, r_3, a_2]$ translates into problem over some smaller dictionary $\widetilde{D_{\mathcal{S}}}^t$ and $\widetilde{D_{\mathcal{E}}}^t$ cut down from $D_{\mathcal{S}}$ and $D_{\mathcal{E}}$ by removing some of its columns, i.e.

$$d_{\mathcal{S}}^i \in \widetilde{D_{\mathcal{S}}}^t \Leftrightarrow \theta_{\mathcal{S}}\left(D_{\mathcal{S}} x^{t-1}\right) - \theta_{\mathcal{S}}\left(D_{\mathcal{S}}\left[x\right]_i\right) \in \mathcal{C}_{\mathcal{S}} \tag{9}$$

$$d_{\mathcal{E}}^i \in \widetilde{D_{\mathcal{E}}}^t \Leftrightarrow \theta_{\mathcal{S}}\left(D_{\mathcal{E}} y^{t-1}\right) - \theta_{\mathcal{S}}\left(D_{\mathcal{E}}\left[y\right]_i\right) \in \mathcal{C}_{\mathcal{E}}, \tag{10}$$

where $\mathcal{C}_{\mathcal{S}} \ni \mathcal{C}$ and $\mathcal{C}_{\mathcal{E}} \ni \mathcal{C}$ denote convex sets defining an expected range of $[r_1, a_1]$ and $[r_1, r_2, a_2]$, respectively.

Fast numerical method for computing state estimates u^t.

Initialization:
Initialize $u^{(0)}$ and $\mathcal{C}$.
Find $D_{\mathcal{S}}, D_{\mathcal{E}}, B$ using learning techniques from [15]
Main loop:
For $t = 1, 2, \ldots$
- Find $\widetilde{D_{\mathcal{S}}}^t \in D_{\mathcal{S}}, \widetilde{D_{\mathcal{E}}}^t \in D_{\mathcal{E}}$ using (9,10)
- Find $\widetilde{B}^t$as a smaller matrix cut down from B taking into account (9,10)
- Find E^t around tool motion center u^{t-1}
- Find $\widetilde{x}^t, \widetilde{y}^t$ as a solution of (1) using dynamic programming
- Find x^t, y^t as a complement of $\widetilde{x}^t, \widetilde{y}^t$
- Find u^t as a solution of (4)

IV. Experiments

We evaluate our tracker on the task of in-vivo tracking of the center of a single instrument in (i) Retinal Microsurgery (dataset with 3 sequences [26]), and (ii) Laparoscopy (dataset with 1 sequence [24]). Both datasets and tool center annotations are publicly available. We train our structural correlation filters on the first halves of the sequences and test them on the reminaing halves, as in [27]. We manually initialize the state variable $u^{(0)}$. We set interframe changes of scale a_1, a_2 to $\times 0.9$ $\times 1.0$, and $\times 1.1$, of shaft rotation r_1 to $\pm 20^\circ$ and of end-effector rotation r_2, r_3 to $\pm 10^\circ$.

Evaluation metric We compare our tracker to state-of-the-art methods in surgical tool tracking: POSE [27], ITOL [28], DDVT [26], and our previous method from [15]. To this end, we follow the standard evaluation metric of thresholded detections to evaluate the proposed method. Namely, a candidate detection $\widehat{c}_i$ is true positive when it falls within the circle of radius T pixels that is anchored at the ground truth tool center location c_i, such that $\|\widehat{c}_i - c_i\| < T$. We refer to this metric as Keypoint Threshold (KT), after [27].

Quantitative and qualitative results The results presented in Fig. 2 show significant improvement of the proposed tracker over our previous tracking-by-detection method [15]. Our single-thredead, Matlab implementation requires $\sim 0.2s - 1.0s$ to process an image frame while [15] requires dozens of seconds. Moreover, it is either on par or better than other methods. Method ITOL is most competitive with respect to

ISBN 978-0-9928626-7-1 © EURASIP 2017

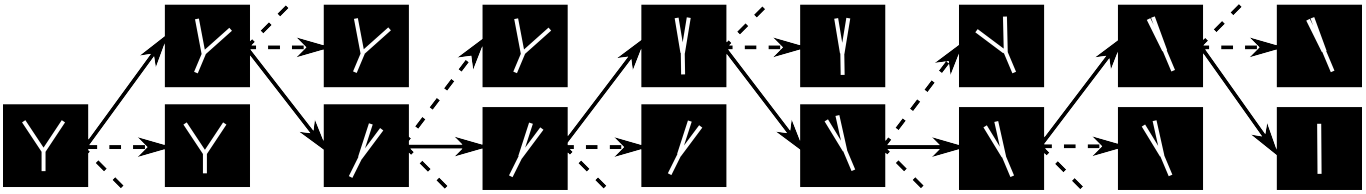

Fig. 1. On-line selection of correlation filters for articulated tool pose estimation. At given time instant and correlation filter, we select a compact subset of correlation filters, which are in a priori defined proximity to the filter. We then match the selected filters to entry image and pick a new filter at given location and scale that best matches current image evidence. As the filters encode pose-specific tool appearance, repeating the procedure generates a tracejctory (solid line) of articulating and rotating tool (here, from Y-open to I-closed forceps). In practice, our experiments show that working already with a small fraction of correlation filters out of a pool of hundreds of filters suffices to produce a stable tool track.

our tracker. Our tracker is better on RM1 but worse on Lapa sequence. However, our method has richer output than ITOL. Apart from tool center, it outputs end-effector articulation and shaft orientation, as in [27]. Additionally, the results presented in Fig. 3 show that our method can successfully track tool pose (e.g., RM 1 and RM 2), recover from lost track (RM 3), while being disrupted by smoke and strong shaft truncations (Lapa).

V. Conclusion

We have described a surgical instrument tracking procedure that achieves state-of-the-art results on two public benchmarks. It combines the strengths of structural, collaborative filtering of dictionaries of discriminative features, generalization properties of SVMs, and Bayesian tracking framework. In particular, we have formulated an appearance model that promotes consistent tool structures by: (a) enforcing shaft parts to be represented by the same template; (b) favouring solutions combining consistent co-appearance of end-effector and shaft poses. Next, we have introduced orientation parameters and scale, which jointly control the dynamics of the target surgical instrument. Despite the inevitable appearance changes of the tracked tool, the proposed scheme allows the tracker to output tool pose and position reliably. In future work, we will extend our work by considering more than two consecutive frames.

Acknowledgment

This work was partly supported by the National Science Centre in the framework of the project number UMO-2014/13/D/ST7/03358.

References

[1] A. Reiter, P.K. Allen, and T. Zhao, "Feature classification for tracking articulated surgical tools.," in *MICCAI 2012, Part II.*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds., vol. 7511, pp. 592–600. Springer, 2012.

[2] S. Voros, E. Orvain, P. Cinquin, and J.-A. Long, "Automatic detection of instruments in laparoscopic images: a first step towards high level command of robotized endoscopic holders," in *The First IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics, 2006. BioRob 2006.*, Feb 2006, pp. 1107–1112.

[3] S. Speidel, G. Sudra, J. Senemaud, M. Drentschew, B. P. Mller-Stich, C. Gutt, and R. Dillmann, "Recognition of risk situations based on endoscopic instrument tracking and knowledge based situation modeling," 2008.

[4] S. Speidel, M. Delles, C. N. Gutt, and R. Dillmann, "Tracking of instruments in minimally invasive surgery for surgical skill analysis.," in *MIAR*, G.-Z. Yang, T. Jiang, D. Shen, L. Gu, and J. Yang, Eds. 2006, vol. 4091 of *Lecture Notes in Computer Science*, pp. 148–155, Springer.

[5] F. Lalys, D. Bouget, L. Riffaud, and P. Jannin, "Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures.," *Int. J. Computer Assisted Radiology and Surgery*, vol. 8, no. 1, pp. 39–49, 2013.

[6] 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*, vol. 35, pp. 633–654, 2017.

[7] T. Liu, G. Wang, and Q. Yang, "Real-time part-based visual tracking via adaptive correlation filters," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4902–4912.

[8] T. Liu, G. Wang, Q. Yang, and L. Wang, "Part-based tracking via discriminative correlation filters," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2016.

[9] M. Zhang, J. Xing, J. Gao, X. Shi, Q. Wang, and W. Hu, "Joint scale-spatial correlation tracking with adaptive rotation estimation," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[10] S. Liu, T. Zhang, X. Cao, and C. Xu, "Structural correlation filter for robust visual tracking," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4312–4320.

[11] Osman Akin, Erkut Erdem, Aykut Erdem, and Krystian Mikolajczyk, "Deformable part-based tracking by coupled global and local correlation filters," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 763 – 774, 2016.

[12] D. Du, H. Qi, L. Wen, Q. Tian, Q. Huang, and S. Lyu, "Geometric hypergraph learning for visual tracking," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–14, 2016.

[13] Fanghui Liu, Tao Zhou, and Jie Yang, "Geometric affine transformation estimation via correlation filter for visual tracking," *Neurocomputing*, vol. 214, pp. 109 – 120, 2016.

[14] L. N. Gaxiola, V. H. Díaz-Ramírez, J. J. Tapia, A. Diaz-Ramirez, and V. Kober, *Robust Face Tracking with Locally-Adaptive Correlation Filtering*, pp. 925–932, Springer International Publishing, Cham, 2014.

[15] D. Wesierski, G. Wojdyga, and A. Jezierska, "Instrument tracking with rigid part mixtures model," in *Computer-Assisted and Robotic Endoscopy: CARE, Held in Conjunction with MICCAI, Munich, Germany, Oct. 2015*, X. Luo, T. Reichl, A. Reiter, and G.-L. Mariottini, Eds., pp. 22–34. Springer International Publishing, 2016.

[16] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M. M. Cheng, S. L. Hicks, and P. H. S. Torr, "Struck: Structured output tracking with kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2096–2109, Oct 2016.

[17] Z. Zhang and K. H. Wong, "Pyramid-based visual tracking using sparsity represented mean transform," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1226–1233.

[18] D. Wesierski, P. Horain, and Z. Kowalczuk, "Ebe: Elastic blob ensemble for coarse human tracking," in *2012 19th IEEE International Conference on Image Processing*, Sept 2012, pp. 1625–1628.

[19] D. Wesierski and P. Horain, "Pose-configurable generic tracking of elongated objects," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 2920–2927.

[20] X. Jia, H. Lu, and M. H. Yang, "Visual tracking via coarse and fine structural local sparse appearance models," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4555–4564, Oct 2016.

[21] T. Zhang, S. Liu, C. Xu, Shuicheng Yan, B. Ghanem, N. Ahuja, and M. H. Yang, "Structural sparse tracking," in *2015 IEEE Conference*

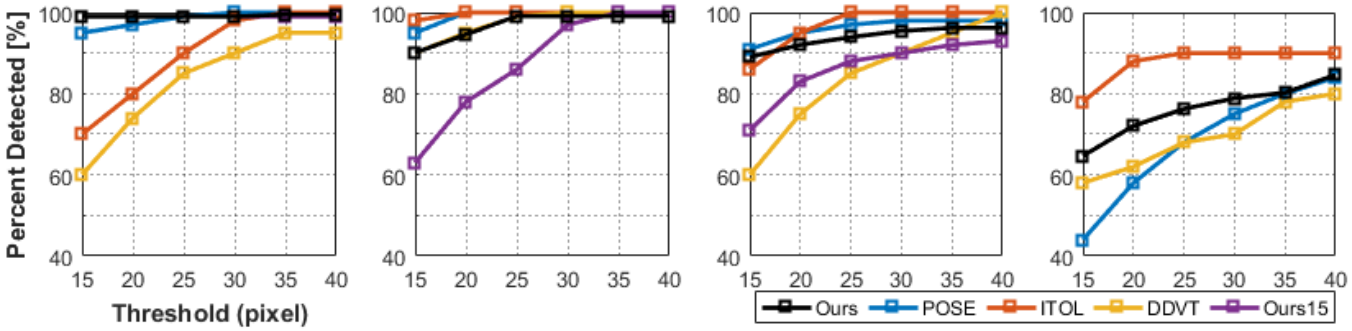

Fig. 2. Quantitative results on Retinal Microsurgery and Laparoscopy datasets (best viewed in color) using KT metric. From left to right: RM1, RM2, RM3, Lapa. Methods: POSE [27], ITOL [28], DDVT [26], Ours15 [15], Ours (black). Our tracker shows significant improvement over our previous tracking-by-detection method (green). Moreover, it is either on par or better than other methods. Method ITOL is most competitive with respect to our tracker. Our tracker is better on RM1 but worse on Lapa sequence. However, our method has richer output than ITOL. Apart from tool center, it outputs end-effector articulation and shaft orientation.

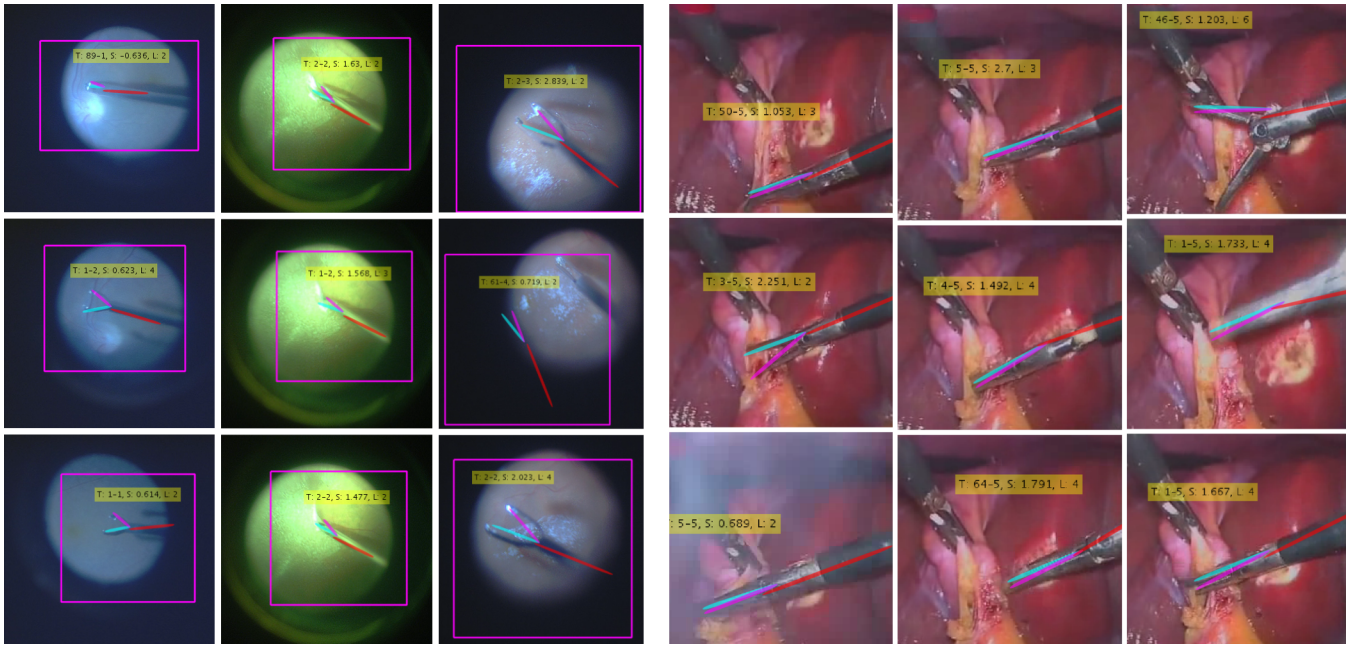

Fig. 3. Qualitative results on Retinal Microsurgery (left) and Laparoscopy (right) datasets. Instead of searching for the tool in the whole image frame, our method reduces the search space over location, scale, and appearance templates based on the detected tool pose in the previous frame. The tool pose serves to initiate new bounding box around the whole tool as well as new appearance templates which are in proximity to the detected end-effector and shaft templates.

on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 150–158.

[22] A. Rodriguez, V. N. Boddeti, B. V. K. V. Kumar, and A. Mahalanobis, "Maximum margin correlation filter: A new approach for localization and classification," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 631–643, Feb 2013.

[23] J. Zhou and P. Shahram, "Visual tracking of laparoscopic instruments," *Journal of Automation and Control Engineering*, vol. 2, no. 3, pp. 234–241, Sept. 2014.

[24] R. Sznitman, C. Becker, and P. Fua, "Fast part-based classification for instrument detection in minimally invasive surgery," in *MICCAI 2014, Part II. LNCS*, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds., vol. 8674, pp. 692–699. Springer, 2014.

[25] G. Demoment, "Image reconstruction and restoration: overview of common estimation structures and problems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 2024–2036, Dec 1989.

[26] R. Sznitman, K. Ali, R. Richa, R.H. Taylor, G.D. Hager, and P. Fua, "Data-driven visual tracking in retinal microsurgery," in *MICCAI 2012, Part II. LNCS*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds., vol. 7511, pp. 568–575. Springer, 2012.

[27] Nicola Rieke, David Joseph Tan, Chiara Amat di San Filippo, Federico Tombari, Mohamed Alsheakhali, Vasileios Belagiannis, Abouzar Eslami, and Nassir Navab, "Real-time localization of articulated surgical instruments in retinal microsurgery," *Medical image analysis*, vol. 34, pp. 82–100, 2016.

[28] Y. Li, C. Chen, X. Huang, and J. Huang, "Instrument tracking via online learning in retinal microsurgery.," in *MICCAI 2014, Part I. LNCS*, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds., vol. 8673, pp. 464–471. Springer, 2014.

ISBN 978-0-9928626-7-1 © EURASIP 2017