Instrument Tracking with Rigid Part Mixtures Model

Daniel Wesierski[1(✉)], Grzegorz Wojdyga[1], and Anna Jezierska[2]

[1] Multimedia Systems Department, Faculty of Electronics, Telecommunications, and Informatics, Gdansk University of Technology, Gdańsk, Poland
daniel.wesierski@pg.gda.pl
[2] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Abstract. Tracking instruments in video-assisted minimally invasive surgeries is an attractive and open computer vision problem. A tracker successfully locating instruments would immediately find applications in manual and robotic interventions in the operating theater. We describe a tracking method that uses a rigidly structured model of instrument parts. The rigidly composed parts encode diverse, pose-specific appearance mixtures of the tool. This rigid part mixtures model then jointly explains the evolving structure of the tool parts by switching between mixture components during tracking. We evaluate our approach on publicly available datasets of *in-vivo* sequences and demonstrate state-of-the-art results.

Keywords: Instrument tracking · Video-assisted minimally invasive surgery · Part-based models

1 Introduction

Locating instruments in videos for augmented assistance [1] during minimally invasive surgeries (MIS) has recently received much attention. Minimally invasive surgeries offer a number of advantages over open surgeries. Less postoperative pain, reduced blood loss, minor scarring, and shorter recovery time and hospitalization are attractive factors for inpatients and clinicians. Carrying out such a surgery, though, is a challenging task. The surgeons first make keyhole incisions in the body to insert elongated surgical instruments. Confronted with lost vision and hampered dexterity, the surgeons require additional sensing devices to monitor the instruments maneuvering within the body. While robotic manipulators can control the instruments with high flexibility and stability, their encoders accumulate errors in forward kinematics and lead to inaccurate estimations of the absolute instrument location [11]. On the other hand, specialized hardware sensors and encoders require extensive hardware integration and suffer from lower accuracy [2] thereby cumbersomely integrating to multiple operating rooms. Arguably, widespread color cameras in MIS offer a natural, visual feedback to surgeons. Other imaging modalities such as depth-only sensing devices would be

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 22–34, 2016.
DOI: 10.1007/978-3-319-29965-5_3

hardly interpretable. Amenable to easy transfer between operating rooms and motivated by steady progress of computer vision, vision-based instrument tracking thus constitutes an encouraging approach to improving the guidance and navigation of manual and robotic surgeries.

Description of image features plays a significant role in MIS tool tracking setting. Registered videos may suffer from degraded quality, e.g., due to motion blur. Moreover, adverse lighting conditions in the form of globally varying illumination of the scene, specular reflections on the tool and tissue regions, as well as shadows left by the tool are factors that make tool detection a challenging task in practice. Past work has explored color and gradient features [2,11] to discern greyish tool foreground from reddish and whitish tissue background, markers [13], and used elaborate classification schemes during detection [4,5]. Bootstrapping object appearance from initial frame, that reported remarkable results in the general object tracking setting [16], has recently also been applied to tracking MIS instruments [10] with state-of-the-art performance.

We describe a rigid part mixtures model of a surgical instrument and a detection procedure for tracking its 2D pose (i.e., center and orientation) in MIS videos. As the 3D pose can be recovered from stereo-cameras [1], here we focus on the problem of 2D pose estimation in a single image. While motion models can be used for filtering of, e.g., instrument location and size [9], we achieve good tracks by detecting the instrument pose in each frame independently from neighbor frames. Our model is a spatial assembly of instrument parts that encode mixtures of dedicated pose appearances. By capturing such appearances of an object part at various poses, our approach relates to poselets [15] that reason about fragmented object pose from rigid parts. It differs from poselets by jointly modeling the compositions of small and large part mixtures that can explain full pose of the instrument. Consequently, our approach leverages successful flexible part mixtures model [6] that can be trained with datasets of modest size [17]. Structured part-based models use deformation constraints that act like springs to flex the model to regions with putative objects. Arranged under a tree-graph, they can efficiently explain previously unseen configurations of the flexible object structure but, at the same time, such models can overlap two tip parts on one tip of the tool. In the spirit of poselets, we avoid double-counting image evidence [18] by rigidly modeling end-effector articulations with larger, rigid parts. Hence, our approach differs from past work by enforcing strictly rigid, global compositions of part mixtures and by consistently capturing variable instrument structure.

Our contribution is two-fold. Firstly, we develop a springs-free, structured part-based model of an instrument. It imposes a rigid structure on spatially distributed local features to discard putative tool regions, e.g. in tool neighborhood, that might prompt models with springs to incorrect or flexed structure detections. Secondly, we demonstrate that a structured part-based model can be successfully applied to instrument tracking in MIS. Estimating instrument pose is typically approached in a disjoint manner by first detecting individual parts and then fusing detections with, e.g., a Kalman filter [1] or RANSAC sampling [5]. By exploiting rigidly structured relations between instrument parts,

our tracker detects the end-effector and shaft parts jointly thereby recovering the instrument pose. Applying a structured model, though, is challenging as this requires frequent updates of its underlying structure. Object appearance can vary significantly between frames, especially due to frequent truncations. Specifically, the rigid, straightly elongated shaft has often been used as a discriminative visual cue in detecting the tool and in estimating its 3D pose [3,14]. However, observing that surgeons often prefer to work in close proximity to tissue, [12] ignore the shaft and focus on tracking the articulating end-effector with thousands of efficiently matched templates.

This leads to a dilemma. On the one hand, one would like to take advantage of the shaft part when it is visible. On the other hand, one has to take into consideration the varying, truncated tool structure. Our model-based tracker exploits the rigid shaft while adapting to its changing length. We then discriminatively train dedicated models on a series of training images for each video sequence and show that our method is on par with or exceeds state-of-the-art results in instrument tracking on publicly available datasets.

2 Method

The structure of MIS instruments, e.g. for laparoscopy, retinal microsurgery, in image I can operationally be represented as a pair composition of two parts: (i) a rigid, straight, elongated shaft $\mathcal{S}$ and (ii) a rigid or an articulated end-effector $\mathcal{E}$, as depicted in Fig. 1. Let G_I denote a two-dimensional regular tessellation of the pixel grid of image I, $L \subset \mathbb{N}^{1\times 2}$ a discrete set of locations on the whole grid G_I, and $L_b \subset L$ a discrete set of locations on an arbitrarily shaped (e.g., rectangular, circular) one-dimensional border stripe of this grid.

The $\mathcal{E}$-part is enclosed in a single window in the grid with the center location $l_{\mathcal{E}} \in L \setminus L_b$. An $\mathcal{S}$-part is a collection of N_e subparts that are outlined by adjacent windows. We restrict possible locations of these windows $l_{\mathcal{S}(k)} \in L$, where $1 \leqslant k \leqslant N_e$, to an oriented raster line segment[1] that anchors at $l_{\mathcal{E}}$ and ranges from the border stripe $l_{\mathcal{S}(1)} \in L_b$ to the end of the shaft $l_{\mathcal{S}(N_e)} \in L$. Then, let $l = (l_{\mathcal{E}}, l_{\mathcal{S}(1)})_{1\times 4}$ denote the line segment. As the shaft is often truncated and partially occluded, we represent the $\mathcal{S}$-part as a subcollection of $N \leqslant N_e$ subparts for each new image frame I. As a result, in our model the location of the $\mathcal{S}$-part $l_{\mathcal{S}}(l) = \big(l_{\mathcal{S}(k_1)}(l), \ldots, l_{\mathcal{S}(k_N)}(l)\big)_{1\times 2N}$ determines some ordering of these subparts on the line segment l of the instrument.

In practice, both parts slightly rotate during a surgery while instrument pose admits *non-circumvolving* motion. In general, though, the shaft is oriented at an arbitrary angle as the locations of body incisions vary between surgical scenarios. Moreover, the grippers of the end-effector articulate and take various forms, i.e. the length and shape of the grippers varies. In view of this, we approach the

[1] The location of each subpart is $l_{\mathcal{S}(k)} = l_{\mathcal{S}(1)} + \big[s_k \left(l_{\mathcal{E}} - l_{\mathcal{S}(1)}\right)\big]$, where $[\cdot]$ is the nearest integer function, s_k is a scaling factor $0 \leqslant s_k \leqslant s_{N_e} < 1$, and s_{N_e} ensures that the location $l_{\mathcal{S}(N_e)}$ of the window of the last subpart of the $\mathcal{S}$-part does not overlap with the window of the $\mathcal{E}$-part.

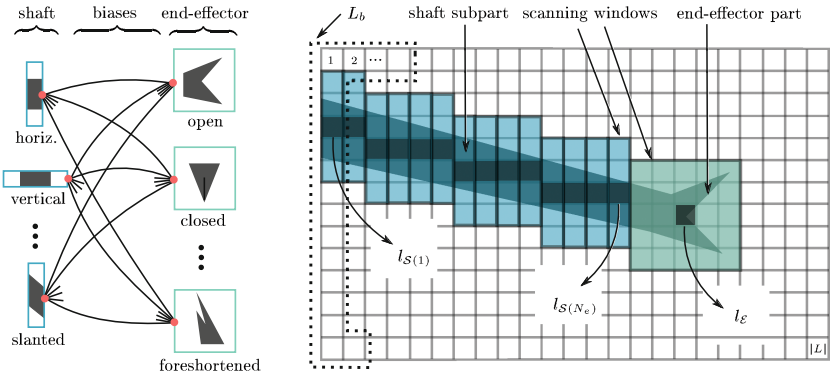

Fig. 1. Our rigid part mixtures model (left) and its instantiation on the grid G_I (right). The model uses (i) a set of appearance templates (i.e., part mixture) that represent a single subpart of the shaft at multiple orientations, (ii) a set of appearance templates that represent various articulations of the end-effector part (e.g., rotated, open or closed gripper), and (iii) a set of biases that promote or discourage rigid appearance compositions of mixture components of the shaft and end-effector parts.

problem of tracking 2D instrument pose by capturing the appearance variation of the tool with a structured model of rigid mixtures of parts that jointly encodes pose-dependent tool appearance.

Model. We represent the appearance and structure of the instruments under graph $M = \{V, E\}$. The appearance mixtures of the end-effector part are chained with the appearance mixtures of the shaft parts. The nodes $V = \{w^i_{\mathcal{E}}, l_{\mathcal{E}}\}^{n_{\mathcal{E}}}_{i=1} \cup \{w^j_{\mathcal{S}}, l_{\mathcal{S}}\}^{n_{\mathcal{S}}}_{j=1}$ denote particular appearances of the $n_{\mathcal{E}}$ end-effector and $n_{\mathcal{S}}$ shaft mixtures, respectively. The i-th component of the appearance mixture of the end-effector part at location $l_{\mathcal{E}}$ is specified by template $w^i_{\mathcal{E}}$ that rigidly encodes specific articulation of this part, as encountered in poselets-based approaches to object recognition [15] and in MIS tool tracking scenarios [12]. The j-th component of the appearance mixture of the shaft part at location $l_{\mathcal{S}}$ is specified by template $w^j_{\mathcal{S}}$ that can capture specific perspective and orientation of the part, e.g. an outwards slanted shaft. The edges $E = \{b^{ij}_{\mathcal{ES}}\}^{n_{\mathcal{E}} \times n_{\mathcal{S}}}_{ij=1}$ model rigid compositions of the end-effector mixture with the shaft mixture. Specifically, the scalar-valued co-occurrences $b^{ij}_{\mathcal{ES}}$ bias configurations of mixtures such that certain, rigidly encoded articulations $w^i_{\mathcal{E}}$ may form more consistent compositions with certain orientations $w^j_{\mathcal{S}}$. In effect, our model admits a strictly rigid structure.

We define the mixture of the shaft part as orientation templates. On the other hand, the $\mathcal{S}$-part lies on the oriented line segment l. Hence, the mapping $j : l_{1\times 4} \rightarrow \mathbb{N}^1$ of a given instance of this oriented line uniquely determines the j-th mixture component of the shaft. Then, instantiating a composition of particular mixture components of the $\mathcal{ES}$-parts in image I at location $l = (l_{\mathcal{E}}, l_{\mathcal{S}(1)})$ is scored with our model as:

$$S(I,l,i) = w^i_{\mathcal{E}}\phi^i_{\mathcal{E}}(I,l_{\mathcal{E}}) + \sum_{p=1}^{N} w^{j(l)}_{\mathcal{S}}\phi^{j(l)}_{\mathcal{S}}(I,l_{\mathcal{S}(k_p)}(l)) + b^{ij(l)}_{\mathcal{ES}} \quad (1)$$

where $\phi^i_{\mathcal{E}}(I,l_{\mathcal{E}})$ and $\phi^{j(l)}_{\mathcal{S}}(I,l_{\mathcal{S}(k_p)})$ are image descriptors (e.g., a `HOG` [8], a color histogram) in the window of the i-th mixture component of the $\mathcal{E}$-part at $l_{\mathcal{E}}$ and in the window of the subpart of the j-th mixture component of the $\mathcal{S}$-part at $l_{\mathcal{S}(k_p)}$, respectively.

The varying length of the shaft notwithstanding, our model allows for taking advantage of the discriminative evidence for this part in each image during tracking. We achieve this with N subparts of the shaft that are anchored at $l_{\mathcal{S}}$. As the elongated shaft roughly admits consistent appearance along the image plane, we deem all subparts of its j-th mixture component to be alike and dedicate a single, canonical template $w^{j(l)}_{\mathcal{S}}$ for representing their appearance. In effect, the subparts, which share the single template, render our model less complex in learning from and matching to images.

Detection. We cast the problem of instrument tracking within the tracking-by-detection framework. We infer the rigid composition of mixture components of the $\mathcal{ES}$-parts at location l that best explains current video frame I by solving the inference problem $\texttt{argmax}_{l,i}\ \ S(I,l,i)$, as depicted in Fig. 2.

Matching the appearance templates $\{w^i_{\mathcal{E}}\}_{i=1}^{n_{\mathcal{E}}}$ and $\{w^j_{\mathcal{S}}\}_{j=1}^{n_{\mathcal{S}}}$ to corresponding image descriptors at each location in L amounts to the convolution in the feature space[2] that yields tables of appearance scores for each mixture component. As our graph M is a mixture of chains, in which $\mathcal{E}$-part mixtures are parents and $\mathcal{S}$-part mixtures are children, we employ dynamic programming as an exhaustive search algorithm over the state space (l,i) to combine the appearance scores across plausible locations and mixture components.

To this end, the search procedure commences by partitioning the border stripe L_b of the grid G_I into $n_{\mathcal{S}}$ disjoint segments $L_b = \bigsqcup_{j=1}^{n_{\mathcal{S}}} L^j_b(l_{\mathcal{E}})$ at given $l_{\mathcal{E}}$. All pairs $(l_{\mathcal{E}}, l_{\mathcal{S}(1)} \in L^j_b(l_{\mathcal{E}}))$ together determine a pencil of line segments. The segments, in turn, indicate all possible orientations of the $\mathcal{S}$-part at $l_{\mathcal{E}}$ within the angular range of the j-th mixture component. As the $\mathcal{S}$-part is represented by N subparts, the score of each hypothesized orientation of the shaft depends on finding such a configuration $l_{\mathcal{S}}(l)$ of image descriptors that best match to the $w^{j(l)}_{\mathcal{S}}$ template. This results in selecting N-best scoring subparts of the shaft within the given line segment.

The search proceeds by enumerating all possible compositions of mixture components of the $\mathcal{ES}$-parts. After aggregating the score $b^{ij(l)}_{\mathcal{ES}}$ of each composition with the N-best scores of the shaft part, the best location $l_{\mathcal{S}(1)}$ of the shaft is selected at given $l_{\mathcal{E}}$ for each i-th mixture component of the end-effector. We then retrieve the best i-th mixture component at $l_{\mathcal{E}}$.

[2] Since the target object can change its scale during tracking, we search over the feature pyramid of $\phi(I,\cdot)$ at run-time.

After repeating this search procedure for each $l_{\mathcal{E}}$, we select $l_{\mathcal{E}}$ with the best aggregated score (1), then backtrack to the best i-th mixture component stored at that location, and terminate at the best $l_{\mathcal{S}(1)}$ pointed by this component.

Learning. We learn the parameters of our rigid part mixtures model in the supervised manner. Our instrument model uses a mixture of appearance templates per part, where only a single template of this part is present in a given positive training image. As we assume a given collection of positive training images contains only keypoint annotations, we retrieve the missing ij-labels of mixture components for each image based on these annotations, as shown in Fig. 3.

We automatically obtain a mixture label of the $\mathcal{E}$ part by first (i) binning the manually labeled end-effector keypoints in a coarse grid (e.g., G_I) and then (ii) grouping the bins features into $n_{\mathcal{E}}$ disjoint sets across all training images. We discard sets with the number of features $< K$. In effect, a given unique spatial arrangement of bins captures a particular articulation of the end-effector. The labels for $\mathcal{S}$-part mixture components are obtained by slicing the image plane into $n_{\mathcal{S}}$ angular intervals. We note when the end-effector part is rigid, we assign the corresponding label of the $\mathcal{S}$-part mixture component to the $\mathcal{E}$-part.

Our rigid part mixtures model is inspired by the flexible part mixtures model [6]. Hence, its array of model parameters is learned jointly and takes the form: $\beta = \left[b^{11}_{\mathcal{ES}}, \ldots, b^{ij}_{\mathcal{ES}}, \ldots, b^{n_{\mathcal{E}} n_{\mathcal{S}}}_{\mathcal{ES}}, w^1_{\mathcal{E}}, \ldots, w^i_{\mathcal{E}}, \ldots, w^{n_{\mathcal{E}}}_{\mathcal{E}}, w^1_{\mathcal{S}}, \ldots, w^j_{\mathcal{S}}, \ldots, w^{n_{\mathcal{S}}}_{\mathcal{S}}\right]$. Since β uses a canonical appearance template $w^j_{\mathcal{S}}$ of a single subpart to *generalize* the appearance of all shaft subparts for j-th mixture component, the function (1) scoring a training feature vector x_n yields the following dot-product form:

$$S(I_n, l, i) = \beta \cdot (0 \ldots 1 \ldots 0 \ldots \phi^i_{\mathcal{E}}(I_n, l_{\mathcal{E}}) \ldots 0 \ldots \phi^{j(l)}_{\mathcal{S}}(I_n, l_{\mathcal{S}(k)}(l)) \ldots 0) = \beta \cdot x_n \tag{2}$$

It induces a sparse structure on x_n that depends on the pre-assignment of mixture labels to respective parts in a given training image I_n.

We then learn the model parameters β with an objective function under linear SVM regime:

$$\begin{aligned} \text{ar gmin}_{\beta,\xi} \quad & \frac{1}{2}\|\beta\|^2 + C^+ \sum_{n=1}^{m^+} \xi_n + C^- \sum_{n=1}^{m^-} \xi_n \\ \text{s.t.} \quad & \beta \cdot x^+_n \geq 1 - \xi_n \ , \quad \forall x^+_n \\ & \beta \cdot x^-_n \leq -1 + \xi_n \ , \quad \forall x^-_n \end{aligned}$$

that can be optimized with, e.g., a dual coordinate-decent solver [6]. The above formulation states that our model β should learn to assign scores higher than 1 to positive examples x^+_n of rigid compositions of respective mixture components and assign scores lower than -1 to negative examples x^-_n. The objective function penalizes violations of these constraints with slack variables $\xi_n \geq 0$, weighted by constants C^+ and C^-. The negative examples x^-_n constitute incorrect detections of the instrument that are mined as hard-negatives on images with masked instruments, as e.g. in [4]. We slightly rotate the positive training images to augment the training set of positive examples.

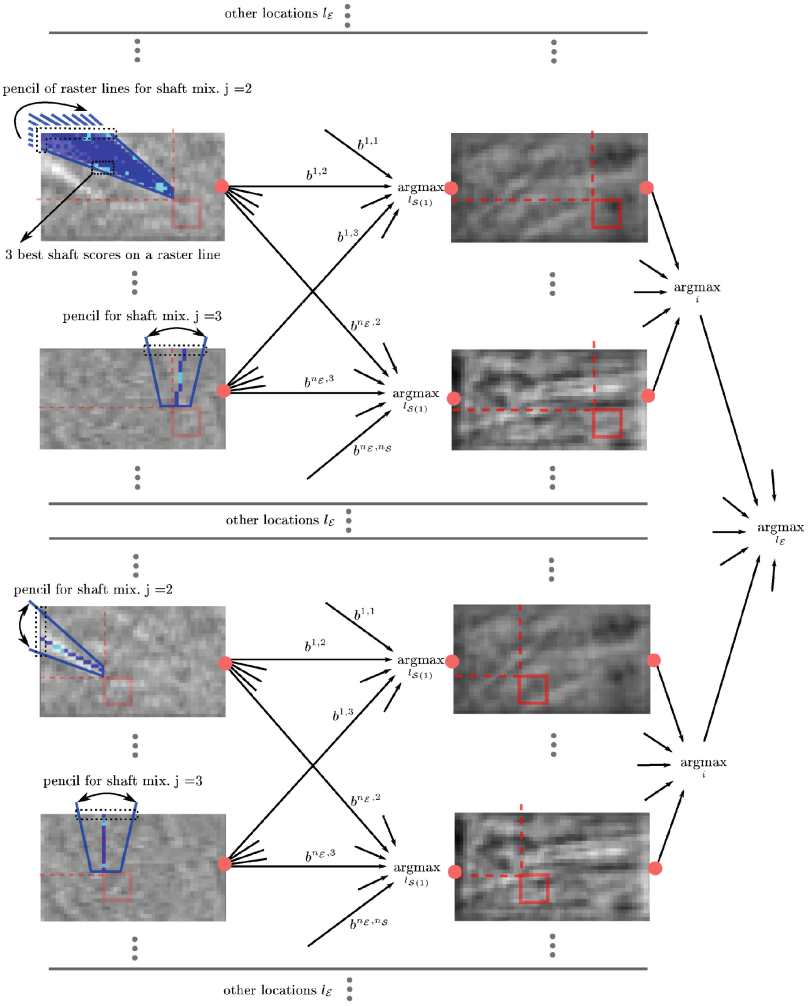

Fig. 2. Instrument detection by matching our model to an image (in feature space). We visualize (best seen in color) two separate iterations. The iterations correspond to two hypothesized locations $l_\mathcal{E}$ of the $\mathcal{E}$-part thereby leading to two different partitions of L_b. In general, for each $l_\mathcal{E} \in L \setminus L_b$, we instantiate $l_{\mathcal{S}(1)}$ yielding all possible oriented line segments that anchor the subparts of the shaft. Each table of appearance scores (in gray) of the $\mathcal{S}$-part (left column) corresponds to j-th mixture component and is selected according to particular instantiation of the line segment. By recursively summing the scores and storing the pointers to selected locations and mixture components, we select (i) N-best scoring subparts of the shaft $l_\mathcal{S}(l)$, followed by (ii) the best line segment $(l_\mathcal{E}, l_{\mathcal{S}(1)})$ per i-th mixture component after adding respective biases, then (iii) best scoring i-th mixture component after adding appearance scores of the $\mathcal{E}$-part (right column), and (iv) terminate by selecting the location $l_\mathcal{E}$ with maximal overall score. As an implementation detail, in the tables the score locations are shifted from the center to upper left corner of every window (Color figure online).

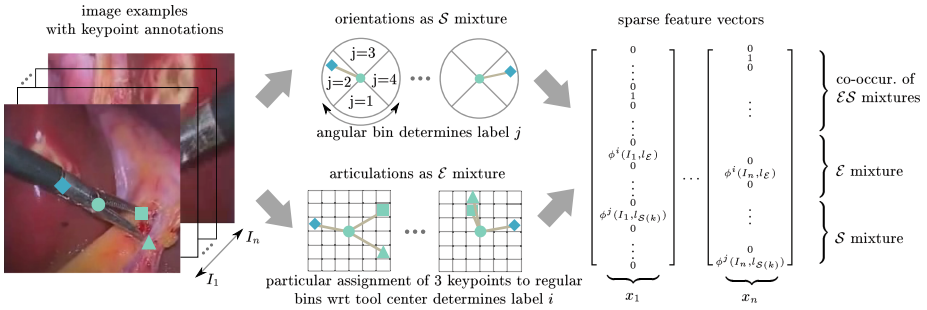

Fig. 3. Learning the mixture labels of the shaft and end-effector parts from image examples determines the sparse structure of feature vectors x_n for SVM classification. The annotations of positive training examples indicate the keypoint locations of two tool tips, tool center $l_\mathcal{E}$, and the end of the shaft. The number of mixture components of both parts, $n_\mathcal{E}$ and $n_\mathcal{S}$, is obtained automatically and depends on the resolution of two respective coarse grids. We use a polar grid to retrieve an orientation type of the shaft part. Then, we rigidly capture articulations of the end-effector part. We first quantize the locations over a regular grid that result in binary occupancy features. We then find their unique groups to retrieve the types of end-effector articulation. Finally, compositions of $\mathcal{ES}$ mixtures serve to store the mixtures co-occurrence indicator and feature descriptors at respective locations in the sparse feature vectors.

3 Results

In this section, we extensively evaluate our method on the task of *in-vivo* single instrument tracking in (i) retinal microsurgery - RM (dataset with 3 sequences [4]), and (ii) spine and pelvic surgery - SPS (dataset with 3 sequences [5]). Both datasets are publicly available.

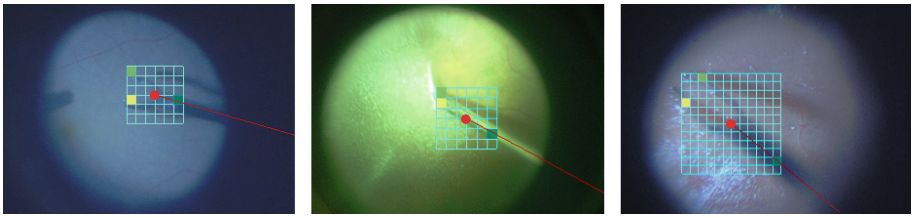

Fig. 4. Instrument pose detection during retinal microsurgery (best viewed in color). Our model detects the tool center (red dot) and the orientation of the shaft (red line). Here, we visualize the windows of end-effector mixture components which are detected on the `HOG` grid (blue). In the spirit of poselets, the model can reason about the articulation of the end-effectors (filled bins on the grid) (Color figure online).

Table 1. Results of our method wrt [5] for estimating the tool center and tool orientation in sequences from RM and SPS datasets. We use the protocol of [5] and evaluate the performance based on the angular mean (Ang. M.) and angular standard deviation (Ang. St.D.) (left) and mean distance (M.) and standard deviation (St.D.) (right). The ratios (train/test) indicate the number of images used for training and testing the model, after the usage guidelines of the SPS dataset. Following the evaluation protocol that omits false negatives and false positives when the tool is present and absent in test images, respectively, we only evaluate the test images that contain the instrument. We note, though, that our tracker could run for images without the tool as the detection threshold is learned within SVM margins. Best results are indicated in bold.

		Ret. 1	Ret. 2	Ret. 3
# Images	[5]	200/152	-	200/297
	Ours	198/152	100/121	196/287
Ang. M.	[5]	4.18	-	**5.31**
	Ours	**3.42**	3.62	5.66
Ang. St.D.	[5]	3.9	-	4.9
	Ours	**2.37**	1.91	**4.53**

		Pelvic 1	Pelvic 2	Spine
# Images	[5]	400/400	100/490	150/322
	Ours	76/213	337/339	91/247
M.	[5]	24.32	16.15	**3.98**
	Ours	**13.87**	**9.07**	13.09
St.D.	[5]	**15.21**	**10.8**	**1.75**
	Ours	45.75	30.25	33.72

Implementation Details. For all sequences, we *equally* configure our model and use *fixed* parameter settings. To make the comparison fair, we follow [4,5] and use the training set of each sequence, as specified in the datasets, to train dedicated models of the instruments. We compute window sizes of the end-effector and shaft parts from the keypoint annotations in the training images.

The appearance templates are defined in `HOG` feature space [7]. We set `sbin`= 8 for `HOG` cells, $K = 10$ for pruning groups of bins features when learning the $\mathcal{E}$ mixtures, and $N = 3$ for the number of detected shaft subparts. To specify the orientation labels for the $\mathcal{S}$-part, we follow the `HOG` specification of 18 equal orientation intervals over $(-\pi, +\pi\rangle$. The number of labels $n_{\mathcal{S}}$ is then determined based on the annotated instruments in the training set. We set $C^+ = 0.004$ and $C^- = 0.002$ to account for $m^+ < m^-$ imbalance in the training set.

Qualitative Evaluation. We qualitatively show that our model can detect the 2D pose of the instrument (i.e., center location of the end-effector and orientation of the shaft) as well as the articulation of the end-effector, as shown in Fig. 4. In RM sequences, the tracker yields robust tracks despite illumination variations and disrupting tool-like shadows. In SPS sequences, the instruments significantly change their scale, are partially occluded, and often heavily truncated. Our method is able to successfully locate the end-effectors in these sequences. It can adapt to the varying length of the shaft by searching for the best-scoring subparts along the hypothesized, oriented shafts (Fig. 5).

Quantitative Evaluation. We report quantitative results in Table 1 as mean distance precision and standard deviation from the ground truth (i) tool center for SPS and (ii) tool orientation for RM. In addition, we report the percentage

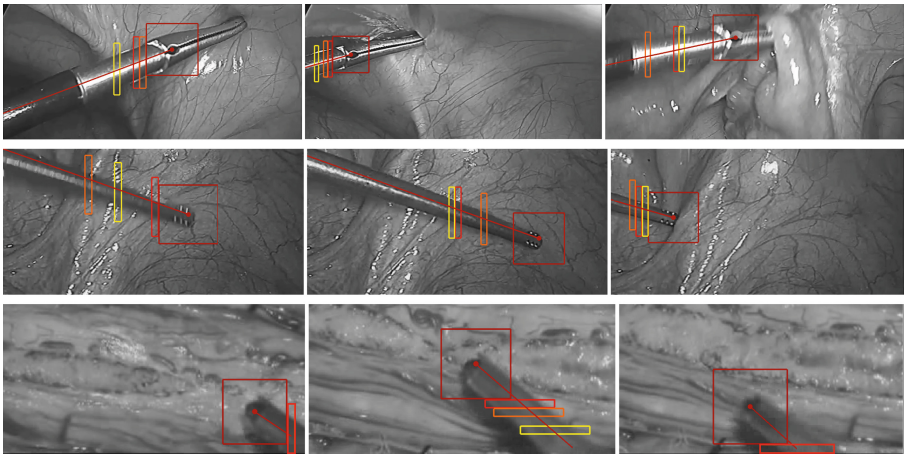

Fig. 5. Instrument pose detection (best viewed in color) in pelvic 1 (top row), pelvic 2 (middle), and spine (bottom) sequences. We show (i) the windows of 3 best shaft subparts that are detected with a single, canonical template and (ii) the window and its center of the end-effector part. We learn the appearance of the $\mathcal{E}$-part based on its window center as indicated by the SPS dataset annotations. The end-effector and shaft mixture labels are equal as the $\mathcal{E}$-part is considered rigid having no articulations in the sequences. Note how mixture components of the shaft part switch to explain the varying orientation of the instrument. Also, the colors red–orange–yellow of the shaft subparts indicate the 1^{st}, 2^{nd}, and 3^{rd} best detection, respectively. The tracker detects the tool at multiple scales (top row). By selecting the best scoring subparts along the shaft, the tracker takes advantage of the discriminative appearance of the shaft (middle row) while at the same time it copes with heavy truncations (bottom row) (Color figure online).

of accurate detections of the tool center within a given pixel range for RM in Fig. 6. We demonstrate that the proposed rigid part mixtures model achieves state-of-the-art results on both benchmarks.

In Table 1, we do well in terms of smaller mean distance precision measure. Our high deviation error wrt [5] for SPS sequences comes from far but rare misdetections of the tool center. In general, though, our method yields stable tracks in the RM and SPS sequences.

In Fig. 6, we are on par with other trackers. We outperform other methods in Ret. 1, but do worse in Ret. 2 wrt [10]. However, while [10] successfully tracks the tool center, our tracker also outputs tool orientation (Figs. 4 and 5). Finally, we examine the reliance of our detector on the length of the shaft. We show that our model, augmented with more subparts of the shaft, better stabilizes the detections thereby leading to improved performance (Fig. 6d).

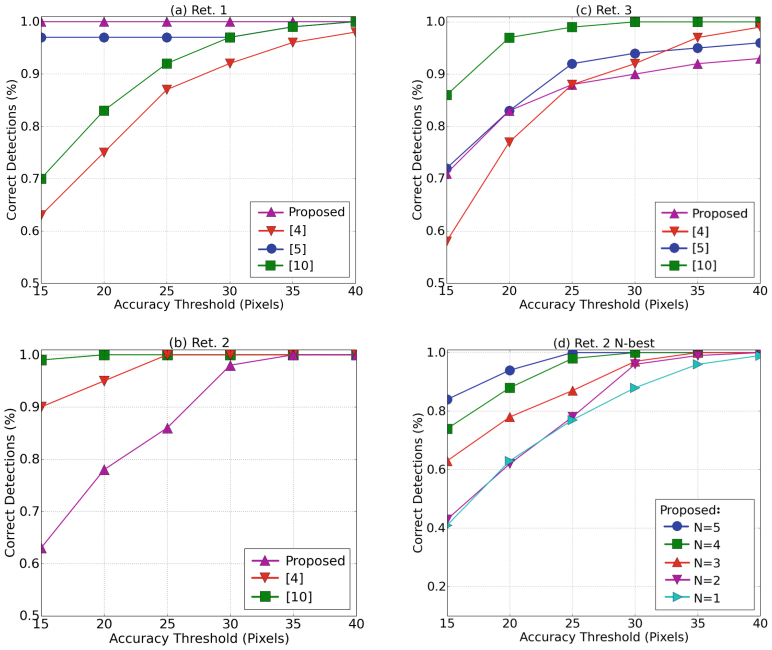

Fig. 6. The results for retina microsurgery dataset on the task of end-effector localization. Our method performs best in Ret. 1 and is on par with [4,5] in Ret. 3. However, we do worse in Ret. 2 and Ret. 3 wrt [10] but additionally output tool orientation. In the last graph (d), we show that the performance of our method scales proportionally with N-best subparts of the shaft. When our model uses 5 subparts, it effectively levels up with [10].

4 Conclusions and Future Work

We proposed a rigid part mixtures model for structurally representing the appearance of surgical instruments in MIS videos. The model robustly explains the evolving object structure by switching between part mixture components that rigidly encode pose-specific appearances of the tool. In effect, our versatile approach to tracking 2D instrument pose reaches state-of-the-art results on two public benchmarks and often improves the estimation of tool location and orientation upon other trackers. We also showed that increasing visual shape cues by a larger pool of shaft subparts leads to more stabilized tool tracking.

Tracking instruments in MIS scenarios is a challenging task. The shaft undergoes frequent truncations, the end-effector can have many degrees of freedom in articulation, such as the da Vinci instruments, and both parts can be occluded when multiple tools are present. At the same time, a tool tracking algorithm should run at frame rates ideally exceeding real-time to minimize the latency of visual feedback and thereby to improve augmented assistance in MIS. Our future work will concentrate on these challenges.

References

1. Reiter, A., Allen, P.K., Zhao, T.: Feature classification for tracking articulated surgical tools. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 592–600. Springer, Heidelberg (2012)
2. Allan, M., Thompson, S., Clarkson, M.J., Ourselin, S., Hawkes, D.J., Kelly, J., Stoyanov, D.: 2D-3D pose tracking of rigid instruments in minimally invasive surgery. In: Stoyanov, D., Collins, D.L., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI 2014. LNCS, vol. 8498, pp. 1–10. Springer, Heidelberg (2014)
3. Wolf, R., Duchateau, J., Cinquin, P., Voros, S.: 3D tracking of laparoscopic instruments using statistical and geometric modeling. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 203–210. Springer, Heidelberg (2011)
4. Sznitman, R., Ali, K., Richa, R., Taylor, R.H., Hager, G.D., Fua, P.: Data-driven visual tracking in retinal microsurgery. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 568–575. Springer, Heidelberg (2012)
5. Sznitman, R., Becker, C., Fua, P.: Fast part-based classification for instrument detection in minimally invasive surgery. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part II. LNCS, vol. 8674, pp. 692–699. Springer, Heidelberg (2014)
6. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2878–2890 (2013)
7. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893. IEEE Press, New York (2005)
9. Sznitman, R., Basu, A., Richa, R., Handa, J., Gehlbach, P., Taylor, R.H., Jedynak, B., Hager, G.D.: Unified detection and tracking in retinal microsurgery. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 1–8. Springer, Heidelberg (2011)
10. Li, Y., Chen, C., Huang, X., Huang, J.: Instrument tracking via online learning in retinal microsurgery. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part I. LNCS, vol. 8673, pp. 464–471. Springer, Heidelberg (2014)
11. Reiter, A., Allen, P.K., Zhao, T.: Appearance learning for 3D tracking of robotic surgical tools. Int. J. Robot. Res. (2013)
12. Reiter, A., Allen, P.K., Zhao, T.: Marker-less articulated surgical tool detection. In: Computer Assisted Radiology and Surgery (2012)
13. Zhao, T., Zhao, W., Halabe, D.J., Hoffman, B.D., Nowlin, W.C.: Fiducial marker design and detection for locating surgical instrument in images. Patent US 068395, 07 08 (2010)
14. Doignon, C., Nageotte, F., de Mathelin, M.: Segmentation and guidance of multiple rigid objects for intra-operative endoscopic vision. In: Vidal, R., Heyden, A., Ma, Y. (eds.) WDV 2005/2006. LNCS, vol. 4358, pp. 314–327. Springer, Heidelberg (2007)
15. Lubomir, B., Malik, J.: Poselets: body part detectors trained using 3D human pose annotations. In: ICCV, pp. 1365–1372. IEEE (2009)

16. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1409–1422 (2012)
17. Zhu, X., Vondrick, C., Ramanan, D., Fowlkes., C.: Do we need more training data or better models for object detection? In: BMVC (2012)
18. Ramakrishna, V., Munoz, D., Hebert, M., Andrew Bagnell, J., Sheikh, Y.: Pose machines: articulated pose estimation via inference machines. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 33–47. Springer, Heidelberg (2014)