Fast Part-Based Classification for Instrument Detection in Minimally Invasive Surgery

Raphael Sznitman, Carlos Becker, and Pascal Fua

École Polytechnique Fédérale de Lausanne, Switzerland
firstname.lastname@epfl.ch

Abstract. Automatic visual detection of instruments in minimally invasive surgery (MIS) can significantly augment the procedure experience for operating clinicians. In this paper, we present a novel technique for detecting surgical instruments by constructing a robust and reliable instrument-part detector. While such detectors are typically slow to use, we introduce a novel early stopping scheme for multiclass ensemble classifiers which acts as a cascade and significantly reduces the computational requirements at test time, ultimately allowing it to run at framerate. We evaluate the effectiveness of our approach on instrument detection in retinal microsurgery and laparoscopic image sequences and demonstrate significant improvements in both accuracy and speed.

1 Introduction

Visual detection and tracking of surgical instruments in minimally invasive surgery (MIS) is an important and challenging problem in computer-assisted intervention. Its goal is to provide accurate 2D or 3D location estimates of surgical instruments from visual data. Critical to a number of applications such as automatic endoscope control [1], instrument-surface detection [2], clinician training evaluation [3] or setting virtual constraints for instrument motion [4,5], instrument detection and tracking can significantly augment the clinicians experience during surgical procedures.

In large part, this detection task is challenging due to illumination changes, specular regions, motion blur and the instrument being partially occluded at times. Among the many approaches proposed in the last twenty years [1,2,6,7,8,9], recent detection-based schemes that rely on building statistical classifiers to evaluate the presence of the instrument appear to be the most promising for *in-vivo* detection and tracking [4,5,10]. Within this last category of methods, Reiter et al. [5] combined a multiclass Random Forest (RF) [11] labelling approach with robot kinematic information to estimate the instrument 3D pose. In [4], RFs were also used to handle instrument-background classification, giving way to instrument segmentations and the 3D pose. Alternatively, [10] combined template tracking and binary classification to provide 2D instrument positions in retinal microsurgery.

However, given that such detection schemes must be fast, these methods usually sacrifice classification accuracy to reduce computational costs at evaluation

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 692–699, 2014.

time. For example, in [4,5], run-time classification is described as a computational bottleneck. As a result it is performed at a single predefined image scale and using only a limited number of RF trees. Furthermore, classification in [5] is only performed on pre-segmented pixels of the instrument, while [10] strongly relies on template tracking to apply a binary classifier on a small window of candidates. Consequently, to attain framerates of 1-5 fps, these methods must ignore structural information about the instrument, such as the position of the shaft, grippers or tip, ultimately impacting detection accuracy.

To overcome these limitations, we propose a robust and efficient 2D instrument detector for MIS procedures based on a multiclass ensemble classifier coupled with an early-stopping algorithm which can deliver both high accuracy and framerate performance. We use gradient boosted regression trees [11] to construct the multiclass classifier to detect different parts of the instrument, as well as the background, and leverage the detected instrument-parts to robustly estimate the instrument pose at arbitrary scales. The early-stopping algorithm extends earlier binary cascades [12,13,14] to the multiclass case. It tracks classification progress for each class using a probabilistic model and evaluates whether further computation is necessary to reliably determine the class of a test sample. In our experiments, we show that our part-based model allows for improved detection accuracy over state-of-the-art methods and can be achieved at framerate speeds due to our early stopping scheme. We validate our framework on *in-vivo* image sequences of microretinal, spine and pelvic surgery.

2 Detection Framework

In this section, we present our instrument detector which assigns an instrument-part or background label to each pixel of an image. We then show how the 2D center and orientation, *i.e.* pose, of the instrument can be estimated from these labels. We will discuss our early-stopping scheme in the following section.

Multiclass Detector. We seek to learn a multiclass classifier from training samples $X = \{x_i, y_i\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ is a gray-scale $r \times r$ image patch with associated class label $y_i \in \{1 \dots C\}$. While choosing which and how many parts should be used is generally a user design choice, Fig. 1 (a) depicts an example where $y_i = 1$ for the background, $y_i = 2$ for the shaft end, $y_i = 3$ for the center and so on. The goal then is to learn a scoring function $F(x) \in \mathbb{R}^C$ which predicts the label of an unseen sample $x \in \mathcal{X}$, by evaluating the function: $Y = \text{argmax}_{c \in \{1 \dots C\}} F^c(x)$, where $F^c(x)$ is the c^{th} dimension of $F(x)$.

While a number of strategies for constructing multiclass classifiers exist, we use the Gradient Boosting framework [11] to learn $F(x)$ by minimizing the squared loss over the training data. Boosting iteratively constructs $F(x)$ as a sum of T computations, or *stages*, such that $F(x) = \sum_{t=1}^{T} f_t(x)$, where $f_t : \mathcal{X} \mapsto \mathbb{R}^C$ is the stage computation at iteration t. To model each stage f_t, we use regression trees of maximum depth 2 [11]. At each regression tree node we evaluate image features which compute the proportions of edges in different locations and

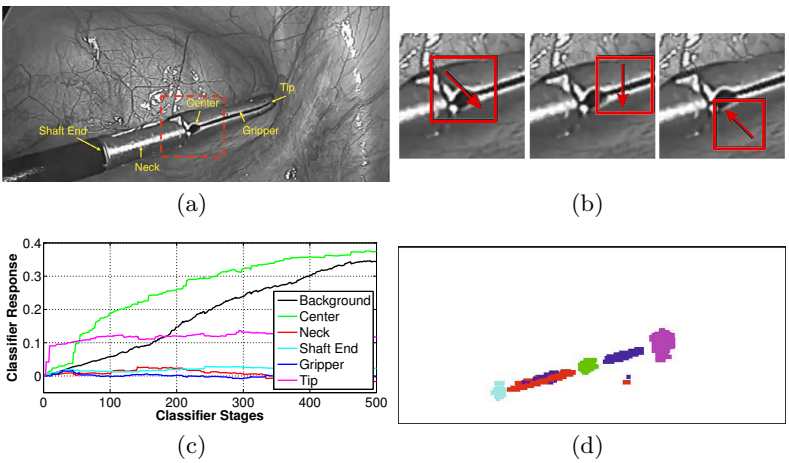

Fig. 1. (a) Example image with instrument-parts highlighted. (b) Examples of three image features overlayed on the evaluated image patch. (c) Plot of accumulated responses as a function of stage evaluations using a trained classifier. Each curve shows the estimate for a specific class. (d) Final response map using the classifier at each image location and scale with non-maximum suppression applied.

orientations of the evaluation window, as in [15]. In Fig. 1(b) we depict three selected features in reference to the center patch from (a) and Fig. 1(c) illustrates the evolution of stage responses of each instrument part after evaluating the patch from (a) using a learned classifier. As in [10], these features can easily be extended to be rotationally invariant and generalize well to variations across image sequences.

Instrument Pose Estimation. To determine the 2D location of the instrument center at test time, we evaluate the classifier $F(x)$ at each image location and at multiple image scales, since the instrument size is a priori not known. Across each scale and position, we then apply non-maximum suppression for each class independently such that non-maxima regions with overlap greater than 99% are suppressed. The maximum score across all classes then determines the label of a pixel. An example of a produced label response map after evaluating the classifier is shown in Fig. 1(d).

From this response map, we perform RANSAC to estimate the overall orientation of the instrument and remove possible response outliers, by fitting a line to the non-maximum suppression output. Weighted averaging [10] on the response scores is then performed for each class to estimate the position of the different parts of the instrument, allowing the instrument center and orientation to be extracted from the part labels. In our experiments, we ran RANSAC with 500 sampling rounds and let inliers be points within 24 pixels of the model, *i.e.* the size of the evaluation patch.

3 Early-Stopping for Multiclass Ensemble Classifiers

Given the trained multiclass classifier described above, our goal is to speed up classification while maintaining accuracy as much as possible. Our key insight is that for most test inputs x, only a reduced number of stages need to be evaluated to estimate with high confidence the class of the test input. We therefore propose the following multiclass early-stopping scheme which extends the ones designed for the binary case [12,13,14] .

First, we consider the class label $Y = \{1, ..., C\}$ to be a discrete random variable with probability distribution $P(Y)$. While this prior is generally task specific, we will assume here and in our experiments that it is uniform for each class label, *i.e.* $\forall y, P(Y = y) = 1/C$. We let $O_{t:t+\delta} = \sum_{m=0}^{\delta} f_{m+t}(x) \in \mathbb{R}^C$ be the sum of stage responses from iteration t to $t + \delta$, where $\delta > 0$, and consider $O_{t:t+\delta}$ to be a random variable with conditional distribution $P(O_{t:t+\delta}|Y = y)$, which we will estimate empirically from data and describe below. We assume that $O_{t:t+\delta}$ is conditionally independent of $O_{j:j+\delta}, j = \{m\delta < t | m \in \mathcal{Z}^+\}$ given the class label and assume $P(O_{t:t+\delta}|Y = y) = \prod_{c=1}^{C} P(O^c_{t:t+\delta}|Y = y)$, where $O^c_{t:t+\delta}$ is the response sum for class c.

With this, we are interested in estimating the likelihood of the label given what has been computed up to any given point. By using these two assumptions, we can thus compute the posterior distribution of Y given $(f_1, \ldots, f_{t+\delta})$ as

$$P(Y = y|f_1, \ldots, f_{t+\delta}) = \frac{1}{Z} P(O_{t:t+\delta}|Y = y) P(Y = y|f_1, \ldots, f_{t-1}), \quad (1)$$

where $Z = \sum_{y'} P(Y = y'|f_1, \ldots, f_{t-1}) P(O_{t:t+\delta}|Y = y')$. In order to estimate the conditional likelihoods $P(O_{t:t+\delta}|Y = y)$ we use a validation set as in [13], *i.e.* labeled data separate from the training data used to train the classifier $F(x)$. To do this, we represent each distribution $\forall (y, c)\ P(O^c_{t:t+\delta}|Y = y)$ using a histogram and use a Parzen window technique with a Gaussian kernel to smooth the estimation.

From Eq. (1) we can compute the posterior distribution of Y at intervals of δ stage evaluations and decide to stop the classification process when the class *uncertainty* is low. We do this by evaluating an approximation to the conditional entropy of the posterior distribution $H(Y|f_1, \ldots, f_t)$ using the Gini Index [11]

1. $t \leftarrow 1$
2. **while** $t \leq T$ and $H(Y = y|f_1, \ldots, f_t) > \gamma$ **do**
3. $\quad O_{t:t+\delta} = \sum_{m=0}^{\delta} f_{t+m}(x)$
4. $\quad$ Compute $P(Y = y|f_1, \ldots, f_{t+\delta})$ using Eq. (1)
5. $\quad t \leftarrow t + \delta$
6. **end while**
7. **return** $\arg_c \max \left\{ \sum_{m=1}^{t} f^c_m(x) \right\}$

Fig. 2. Multiclass Early-Stopping Algorithm

$$H(Y|f_1, \ldots, f_t) \approx 1 - \sum_y P(Y = y|f_1, \ldots, f_t)^2, \quad (2)$$

and stop when its value falls below a threshold $\gamma \in (0, 1)$ specified by the user. Our early stopping algorithm for multiclass ensemble classifiers is summarized in Fig. 2 and has two user parameters: δ, how often the stopping criteria should be evaluated and γ, the entropy threshold which determines when the label uncertainty is sufficiently low. In particular, using large values of γ reduces the overall number of stages evaluated but does so at an accuracy loss since fewer stages are used to estimate test samples. A video demonstrating the functioning of this algorithm is available in the supplementary materials. Note that while this algorithm is used with a boosted classifier in this paper, combining it with RF classifiers as those in [4,5] is also possible.

4 Experiments and Results

We implemented our algorithm in C++ with a detector patch size of 24×24 pixels and evaluated its performance on a MacBook Pro 8-core 2.5Ghz Intel-Core i7. At test time, the detector was evaluated using a 5 pixels stride and a downsampling factor of 1.2. Classification was performed in parallel over pixel locations using the OpenMP library[1].

Retinal Microsurgery. We first evaluated our approach on the publicly available retinal microsurgery instrument dataset [10]. We tested our method on the first and third sequences of the dataset, as the second is extremely short making it ill-suited for training a dedicated detector. For both sequences, we trained our classifier using the first 200 images of the sequences, used the following 50 images for our early stopping algorithm validation set and used the remainder of the images for testing. For the first sequence, we used 4 classes (background, insertion point, tool center and the tool shaft) and for the third sequence we used 3 classes (background, tool center and tool shaft). We set $T = 500$, $\delta = 10$ and $\gamma = 10^{-3}$, and compared our method against [10] and two gradient based trackers [2,8].

As in [10], Fig. 3 depicts our results on this task, where we plotted detection accuracy as a function of the how close the true instrument center is located. In addition, the instrument angle errors for both sequences are on average (and standard deviation) 4.18 (3.9) and 5.31 (4.9) degrees, respectively.

In general our approach is faster and more accurate than previous methods. This is most noticeable in the first sequence, where we significantly outperform [10]. This is in large part because detecting different instrument parts provides a more robust estimate of the pose, while [10] only estimates its center. This is also why our method and [10] perform similarly on the third sequence, as only the instrument center is visible for large segments of this image sequence and thus other instrument parts do not provide additional robustness. The majority of detection mistakes made by our approach are due to motion blur when the instrument moves rapidly. These results can be viewed in the videos submitted in our supplementary materials.

[1] Code and data available at: `http://cvlabwww.epfl.ch/~sznitman/code.html`

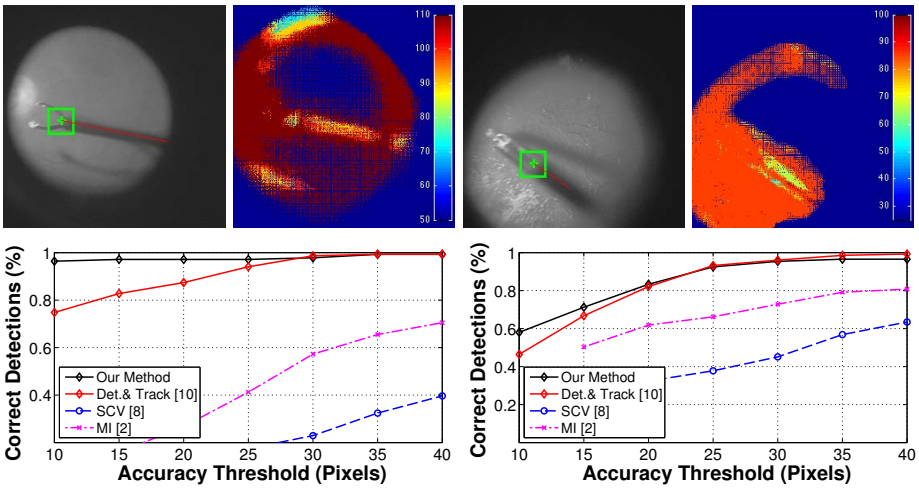

Fig. 3. Examples of retinal microsurgery instrument detections and depiction of the number of stage evaluations at corresponding image regions. For both tested sequences, we plot detection accuracy as function of the detection sensitivity threshold.

Lastly, test images in these sequences are 640×480 pixels large, requiring over 1.5 million classifier evaluations per image. Using the early-stopping technique allows detection to be achieved at 16.6 fps, as roughly only 11 stages are computed on average per evaluation. This is over four times faster than with no early-stopping and roughly twice as in [10]. In Fig. 3, we show the number of stage evaluations at each location on two images, illustrating that different regions are classified with various amounts of computation and that the background is often the most challenging aspect to classify, *i.e.* needs more stages to be evaluated.

Spine and Pelvic Surgery. We also tested our approach on three other *in-vivo* image sequences: two pelvic and one spine procedure. For each sequence, we labeled instrument parts and evaluated our method on unseen consecutive images from the same sequence. Table. 1 summarizes the parameters used to train our detectors and important performance statistics such as the average and standard deviation pixel error of the instrument center estimated with respect to ground truth. Since sequences contain images without instruments in them, we set the training set such that it contains roughly 100 images in it.

In Fig. 4 we show detection results for each sequence, highlighting the different parts detected and overall orientation. Our approach allows various parts of the instrument to be detected as well as the overall orientation, even when some parts are occluded. The complete video results for each sequence are included in the supplementary materials. Most detection errors occur when the instrument rapidly moves into or out of the field of view, inducing significant motion blur,

Table 1. Spine and pelvic surgery image sequence statistics

Seq. Type	# Images train/test	Image size	Parts (# train Samples)	δ / γ	fps	Mean(Std. Dev.) Pix. Error
Pelvic. 1	400 / 400	640×290	Background (10^6), tip (3500) shaft (5000), shaft end (3300) gripper (2500), center (4900)	20 / 50^{-2}	22.5	24.32 (15.21)
Pelvic. 2	100 / 490	640×360	Background (10^6) shaft (5000), center (3500)	20 / 50^{-3}	24.6	16.15 (10.8)
Spine	150 / 322	320×240	Background (10^6) shaft (5000), center (3500)	20 / 50^{-2}	33.1	3.98 (1.75)

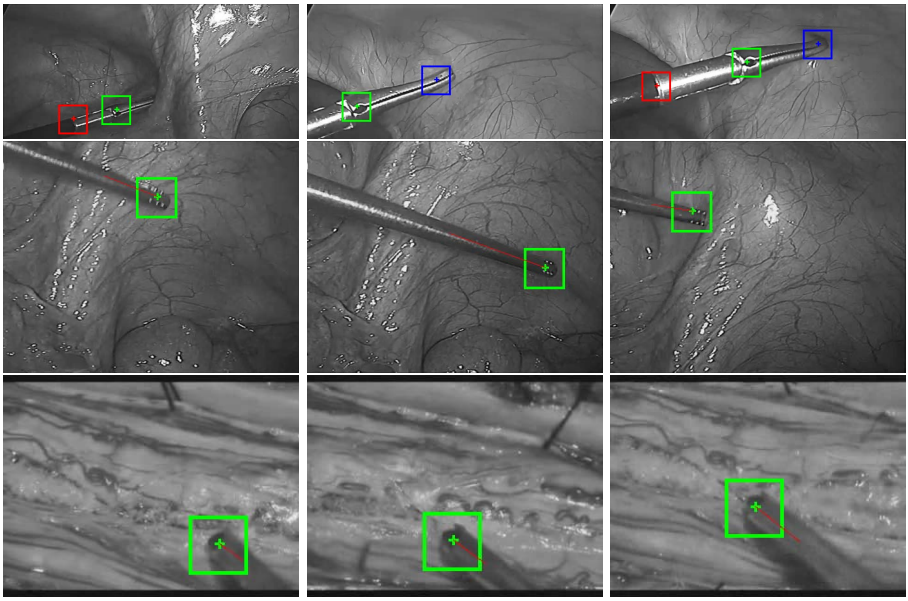

Fig. 4. Visual results on pelvic (rows 1 and 2) and spine (row 3) surgery image sequences. Detected instrument center (green boxes) and orientation (extracted from the detected instrument shaft) overlayed onto the image (best viewed in color).

or when the instrument center is near the image border and little part-evidence can be extracted.

5 Conclusion

We presented a new framework for visual detection of instruments in MIS. Our method constructs an accurate and reliable instrument-part detector from training data, which is then used to estimate the 2D position and orientation of instruments in image sequences. The different instruments parts are key to allowing

the overall instrument pose to be estimated robustly as they provides coherence in the detection results. We also introduced a new early stopping scheme for complete multiclass ensemble classifiers. Our strategy allows a significant reduction in computational requirements at test time, and yields framerate performance. We demonstrated the effectiveness of our approach on spine and pelvic image sequences, outperforming previous methods in both speed and accuracy on microretinal image sequences.

References

1. 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)
2. Richa, R., Balicki, M., Meisner, E., Sznitman, R., Taylor, R., Hager, G.: Visual tracking of surgical tools for proximity detection in retinal surgery. In: Taylor, R.H., Yang, G.-Z. (eds.) IPCAI 2011. LNCS, vol. 6689, pp. 55–66. Springer, Heidelberg (2011)
3. Béjar Haro, B., Zappella, L., Vidal, R.: Surgical gesture classification from video data. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 34–41. Springer, Heidelberg (2012)
4. Allan, M., Ourselin, S., Thompson, S., Hawkes, D., Kelly, J., Stoyanov, D.: Toward detection and localization of instruments in minimally invasive surgery. IEEE Transactons on Biomedical Engineering 60, 1050–1058 (2013)
5. 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)
6. Tonet, O., Ramesh, T., Megali, G., Dario, P.: Trackingendoscopicin- struments without localizer: Image analysis-based approach. Stud. Health Technol. Informat. 119, 544–549 (2006)
7. Burschka, D., et al.: Navigating inner space: 3-d assistance for minimally invasive surgery. Robotics and Autonomous Systems 52, 5–26 (2005)
8. Pickering, M., Muhit, A., Scarwell, J.M., Smith, P.N.: A new multi-modal similarity measure for fast gradient-based 2d-3d image registration. In: Engineering in Medicine and Biology Society, pp. 5821–5824 (2009)
9. Pezzementi, Z., Voros, S., Hager, G.: Articulated object tracking by rendering consistent appearance parts. In: International Conference on Robotics and Automation, pp. 3940–3947 (2009)
10. 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)
11. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2001)
12. Viola, P., Jones, M.: Robust Real-Time Face Detection 57(2), 137–154 (2004)
13. Šochman, J., Matas, J.: Waldboost - Learning for Time Constrained Sequential Detection, pp. 150–157 (June 2005)
14. Sznitman, R., Becker, C., Fleuret, F., Fua, P.: Fast Object Detection with Entropy-Driven Evaluation, pp. 3270–3277 (2013)
15. Fleuret, F., Geman, D.: Stationary Features and Cat Detection 9, 2549–2578 (2008)