Instrument Tracking via Online Learning in Retinal Microsurgery

Yeqing Li[1], Chen Chen[1], Xiaolei Huang[2], and Junzhou Huang[1]

[1] Department of Computer Science and Engineering,
University of Texas at Arlington, Arlington TX 76019, USA
[2] Computer Science and Engineering Department, Lehigh University,
Bethlehem, PA 18015, USA

Abstract. Robust visual tracking of instruments is an important task in retinal microsurgery. In this context, the instruments are subject to a large variety of appearance changes due to illumination and other changes during a procedure, which makes the task very challenging. Most existing methods require collecting a sufficient amount of labelled data and yet perform poorly in handling appearance changes that are unseen in training data. To address these problems, we propose a new approach for robust instrument tracking. Specifically, we adopt an online learning technique that collects appearance samples of instruments on the fly and gradually learns a target-specific detector. Online learning enables the detector to reinforce its model and become more robust over time. The performance of the proposed method has been evaluated on a fully annotated dataset of retinal instruments in in-vivo retinal microsurgery and on a laparoscopy image sequence. In all experimental results, our proposed tracking approach shows superior performance compared to several other state-of-the-art approaches.

1 Introduction

Retinal microsurgery (RM) is an important treatment for sight-threatening conditions. The procedure is performed by a surgeon using a microscope for visualization and manipulating a set of surgical instruments. The operating surgeon faces several difficulties such as indirect visualization of the surgical target, hand tremors and lack of tactile feedback. To overcome these difficulties, new techniques have been developed. Accurate visual tracking of surgical tools in microscopic images is an important technique to complement the previously developed smart tools. In this paper, we focus on the task of robust visual tracking of instruments in in-vivo RM monocular image sequences.

This task is challenging due to the great variability in the appearance of surgical tools because of illumination and other factors. Many existing methods focus on training the appearance model based on color features or the instrument geometry [1–4]. However, these methods often perform poorly under complex appearance changes due to their oversimplified appearance models. Sznitman et al. proposed an approach, namely Data-Driven Visual Tracking (DDVT) [5],

P. Golland et al. (Eds.): MICCAI 2014, Part I, LNCS 8673, pp. 464–471, 2014.

which integrates an instrument detector based on deformable features with a simple gradient-based tracker. DDVT is able to run in video frame rate and achieves state-of-the-art results on challenging human in-vivo surgery datasets. To our best knowledge, DDVT is by far the best visual tracking approach in RM. However, there are two drawbacks to DDVT. First, it needs manually labelled instrument positions in many video frames for training the offline detector. Second, it performs poorly in handling appearance changes that were not observed in the training sequences and could not be modelled by the trained offline detector.

Currently, it draws more and more attentions to integrate online learning techniques in visual tracking system [6, 7]. How to extract new reliable samples without corrupting the current model is a key problem to this kind of systems. Therefore, many techniques have been exploited to constrain the learning process [8, 9]. However, many existing models are not robust enough to apply on RM tracking problem due to the challenges discussed above.

To this end, we propose a new approach based on online learning—Instrument Tracker via Online Learning (ITOL). In this approach, we adopt the paradigm of combining tracking and detection in the same framework [10, 11]. ITOL uses a robust gradient-based tracker capable of failure detection as the basic tracker. Then, a cascade appearance classifier is used as the instrument detector. The appearance model of the detector is initialized by manually clicking the instrument position in the first frame. It is adaptively trained and updated on the fly. Samples for online updating are collected by a filtering process, which selects “unfamiliar” positive samples and “hard” negative samples. The obtained training set is used to augment the model of the detector and prevent the detector from making the similar mistakes. The performance of the proposed approach is evaluated in three human in-vivo retinal microsurgery videos and one laparoscopy image sequence. The experimental results demonstrate that our method significantly outperforms the state-of-the-art approaches.

The rest of this paper is organized as follows: Section 2 introduces the framework and each components of our approach. Then we present our experimental results in Section 3 and conclude the proposed approach in Section 4.

2 Method

In this section, we will detail our proposed method ITOL. Methods for visual tracking usually fall into two groups: tracking through local optimization and tracking by detection [2]. Tracking through local optimization is fast, accurate and able to handle appearance changes of the target. However, continuous template updating is needed in order to maintain accurate position tracking when there are significant changes in target appearance [5]. Tracking by detection has the advantage of being able to handle target disappearance, but the ability of detection is limited by the training data.

Instrument tracking is challenging due to often unexpected appearance changes and extreme deformations of the instrument. We use a multi-component tracking framework to address these problems. A flowchart diagram of the framework is shown in Fig. 1. First, a robust gradient-based **tracker** with the ability of failure detection is used to handle unexpected appearance changes. Then an instrument **detector** is adopted to compensate for tracking loss and it automatically re-initializes the tracker when the instrument reappears after disappearance or tracking loss. To provide more reliable tracking results, outputs of the tracker and the detector will be integrated into a unique target position by a component named **integrator**. Finally, a component named **sample expert** will be used to efficiently select image patches for online updating of appearance model of the detector. In the whole framework, we only need to manually click the position of the instrument in the first frame for training data. Then, the tracking system is fully automatic. Details of each component of the system will be discussed in the following sections.

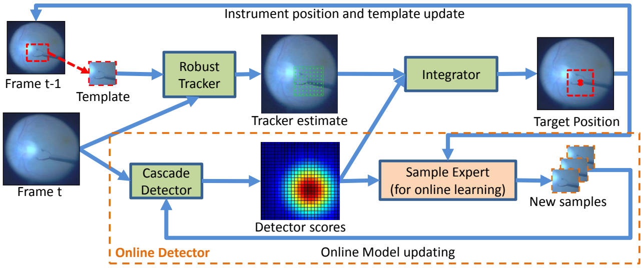


Fig. 1. Diagram of our ITOL framework

2.1 Robust Tracker

The tracker is used to handle instrument appearance changes and bring in new appearance samples. In many cases, although the appearance in the current frame is new to the current model of the detector, it is gradually adapted over time from seen samples. Since we use a gradient based tracker, which is only concerned with similarity between two consecutive frames, it can adaptively collect new appearance samples while tracking. The tracker is based on the Median Flow (MF) algorithm [12]. In the Median Flow tracker, the target is represented by a bounding box around it. For robustness, the bounding box is divided into a $k \times k$ grid ($k = 10$ in our experiments), where each cell of the grid is tracked by the pyramidal L-K algorithm [13]. The displacement of the target is voted by 50% of the most reliable cells. The reliability level of a cell is measured by normalized cross-correlation (NCC).MF also uses a quantity named Forward-Backward (FB) error for failure detection. The tracking is performed both forward and backward along the time axis and the FB error is computed based on the discrepancies between these two trajectories of the target [12]. Since the instrument sometimes move severely or is out of view, this failure detection ability is critical to prevent the tracker from importing false samples.

2.2 Cascade Detector

The gradient-based MF tracker assumes that the target is always in view and under continuous changes. In practice, instruments or tools during RM often undergo large appearance changes, which breaks the assumption. An online detector is developed to compensate for this shortcoming of the tracker and to re-initialize the tracker when an instrument reappears after loss. The detector scans the current frame by sliding window and decides whether the target is present in each window. A complex object detector often requires high computational cost, which makes it impossible for real-time surgical tracking. This problem is addressed by combining successively more complex classifiers in a cascade structure, which rejects most negative windows in the early stages of the cascade thus increasing the processing speed of the detector [14].

In our method, each frame is scanned by the detector at multiple scales using sliding window. All the candidate bounding boxes will be resized to the same size. Inspired by [11], we use a three-stage detector. The first stage is a variance filter that checks if the variance of the patch is under certain threshold related to the variance of trained positive samples. The variance filter can be evaluated efficiently by using integral images [14]. The second stage is random ferns (forest) [15] on patches for comparing the pixel values. Pixels in a patch are first divided into several groups. The probability is then computed for each group based on the number of times that the same feature combination appeared in previous frames as positive or negative examples. The final confidence score is computed by averaging the probabilities of each group. The third stage is a 1-Nearest-Neighbour (1NN) classifier using Normalized Correlation Coefficient (NCC) as the distance between the candidate patch and two sets of patches: positive patches and negative patches. Usually, the first two stages are able to reject more than 95% of the candidate windows, which makes the detector very efficient. In fact, this detector is able to run at nearly 30fps in our experiments.

2.3 Integrator

As discussed above, the detector and the tracker have their respective advantages and disadvantages. Therefore, we use the integrator to integrates their outputs to achieve an optimal estimation. The rules for this integration are: 1) If neither the tracker nor the detector output any positions, the target is declared as not visible; 2) Otherwise, all the outputs of the tracker and the detector are clustered into one by their scores. Suppose s_+ is the similarity between a candidate patch and its nearest neighbour in the positive sample set and s_- is the similarity between the patch and its nearest neighbour in the negative sample set, and $\rho = \frac{s_+}{s_-}$. Then the score of the patch is defined as $s = \frac{1}{1+\rho}$.

2.4 Online Updating of Detector's Model

The sample expert is designed to select new training samples for online model updating of the detector. Online updating make the detector capable of handling unexpected appearance changes and more robust to the noises. Given new samples, the updating process is straightforward. For random ferns, the probability of each branch is updated by adding the results of the pixel comparison. The 1NN classifier simply adds new samples to its sample sets.

The online learning method is detailed in the following. To prevent false positive samples, the sample expert use higher threshold than the detector. Then we consider these bounding boxes as potential positive samples. Starting from the output of the integrator, the sample expert will generate the new positive samples by choosing bounding boxes that are very close to the output one. Second, we filter them by our 1NN classifier and only accept the samples that are rejected by the 1NN classifier. The second step has two effects: 1) It rules out those "easy" samples to avoid redundancy; 2) The remaining samples are "new" enough so that the model will improve very rapidly. In order to accelerate the growth of the model, positive sample are rotated and blurred to generate more data. For negative examples, a common practice is focusing on "hard" samples. Therefore, only samples that have passed the first two stages of the detector and far away from the output are considered candidates of negative samples.

3 Experiment and Results

In this section, we conduct experiments to evaluate ITOL on two public datasets: **Retina Microsurgery Dataset** and **Laparoscopy Sequence** [5].

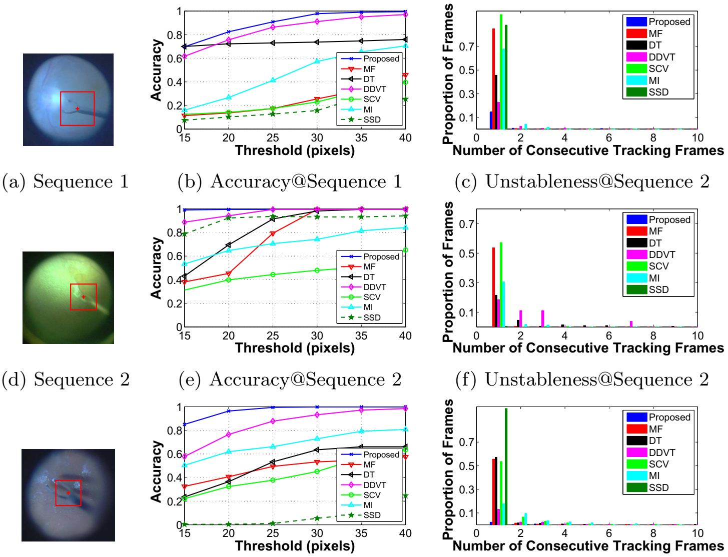
- **Retina Microsurgey Dataset** consists of 3 sequences of in-vivo vitreoretinal surgery, which contains a total of 1171 images (640×480 pixels). See Fig. 2 for examples. These sequences are challenging due to variations in illumination type and quantity, light source position and the presence of blur and shadows.
- **Laparoscopy Sequence** consists of 1000 images with labelled locations of the tool tip. The original video is from Youtube. There are two instruments in each image, hence there are roughly 2000 instrument locations.

We compare our method **ITOL** with four baseline methods: **DDVT** [5], **SCV** [16], **MI** [4], **SSD** [17]. We also compare two components used in the proposed method: **Median-Flow (MF)** and **Detector-Tracker (DT)**. **MF** is the gradient-based tracker that we used. **DT** is MF plus the cascade detector without online model updating. For fair comparison, two measures are used by following the experimental setting of [5]: the accuracy on the thresholding distance to groundtruth and the number of the consecutive tracking frame. The accuracy is defined as the percentage of the detection within δ pixels of the groundtruth annotation. We vary δ from 15 to 40 in experiments (same as the setting in DDVT [5]).

The proposed method is implemented in Matlab. All experiments are conducted on a Desktop PC, 3.4GHz Intel Core i7-3770 and 12GB RAM. Our method runs at nearly 20fps and should run even faster implemented on parallel architecture (e.g. GPU or Mutlti-core).

3.1 Retina Microsurgery Dataset

The experimental results on the RM dataset are shown in Fig. 2. Results of each video sequence are shown in one row. In all the results, DDVT [5] outperforms the others except the proposed ITOL. ITOL also outperforms MF and DT, which validates the benefits of the online detector. Similar trends have been witness in all three videos where ITOL always achieves the best accuracy and unstableness. We accredit the advantages of the proposed ITOL to the online learning component that effectively updates the detector and makes it adapt to the appearance changes of instruments. One thing that is worth to note is DDVT uses the offline detector and therefore requires sufficient amount of training data before tracking (e.g. 500 manually labelled frames [5]), while our method bases on online learning techniques and only requires one labelled position in the first frame as training data before tracking.



(a) Sequence 1 (b) Accuracy@Sequence 1 (c) Unstableness@Sequence 2

(d) Sequence 2 (e) Accuracy@Sequence 2 (f) Unstableness@Sequence 2

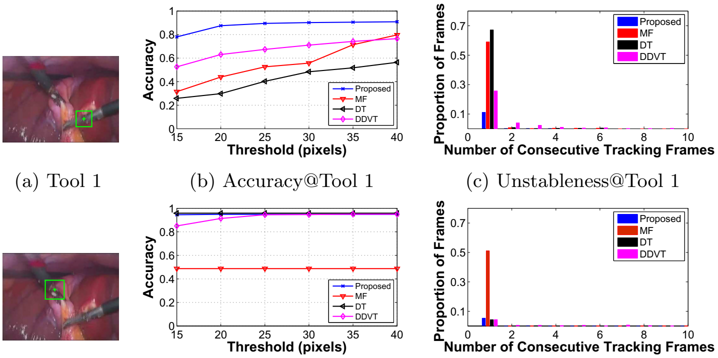
(g) Sequence 3 (h) Accuracy@Sequence 3 (i) Unstableness@Sequence 3

Fig. 2. The results on Retina Microsurgery Dataset. For values of accuracy (the 2nd column), the higher the better. For values of unstableness (the 3rd column), the lower the better.

3.2 Laparoscopy Sequence

Finally, we also evaluate our method on the laparoscopic instrument sequence. The sequence is provided by [5]. DDTV uses the first 500 images for training and the last 500 images for testing. For fair comparison, we follow the setting of [5] and use the last 500 images for testing. However, we only need one image frame for training before tracking because of the online learning technique.

There are two tools in this video. For better visualization, we separately present the experimental results of two instruments in Fig. 3, one in each row. In the sequence, the first tool is under big changes in terms of the instrument structure and movement. Our method significantly outperforms DDVT [5] and two component methods. The second tool is relatively stable in shapes and positions in the whole testing image sequence. The results of the proposed approach are similar to those of the DDVT.



(a) Tool 1 (b) Accuracy@Tool 1 (c) Unstableness@Tool 1

(d)Tool 2 (e) Accuracy@Tool 2 (f) Unstableness@Tool 2

Fig. 3. The results on Laparoscopy Sequence. For values of accuracy (the 2nd column), the higher the better. For values of unstableness (the 3rd column), the lower the better.

4 Conlcusion and Discussion

We proposed a novel approach, dubbed ITOL, for visual tracking of retinal instruments during in-vivo retinal microsurgery. Our method consists of four components: a robust gradient-based tracker, a cascade detector, an integrator and a sample expert. While the first three components make a robust and automatic tracker, the sample expert works to achieve online updating of the appearance model of the detector. ITOL only needs manually labelled position in the first frame and all remaining steps are fully automated, which makes it an approach needing much less user input than other existing methods. ITOL can also automatically re-initialize the tracker after failure. Experimental results on two video datasets demonstrate that the proposed method outperforms the state-of-the-art approaches. Our method makes tracking in RM much more feasible than before.

References

1. Pezzementi, Z., Voros, S., Hager, G.D.: Articulated object tracking by rendering consistent appearance parts. In: IEEE International Conference on Robotics and Automation, ICRA 2009, pp. 3940–3947 (2009)

2. 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)
3. Burschka, D., Corso, J.J., Dewan, M., Lau, W., Li, M., Lin, H., Marayong, P., Ramey, N., Hager, G.D., Hoffman, B., et al.: Navigating inner space: 3-D assistance for minimally invasive surgery. Robotics and Autonomous Systems 52(1), 5–26 (2005)
4. 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)
5. 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)
6. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: BMVC, vol. 1, p. 6 (2006)
7. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE Computer Vision and Pattern Recognition (CVPR), pp. 983–990 (2009)
8. Liu, B., Huang, J., Yang, L., Kulikowsk, C.: Robust tracking using local sparse appearance model and k-selection. In: IEEE Computer Vision and Pattern Recognition (CVPR), pp. 1313–1320 (2011)
9. Liu, B., Yang, L., Huang, J., Meer, P., Gong, L., Kulikowski, C.: Robust and fast collaborative tracking with two stage sparse optimization. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 624–637. Springer, Heidelberg (2010)
10. Sznitman, R., Richa, R., Taylor, R.H., Jedynak, B., Hager, G.D.: Unified detection and tracking of instruments during retinal microsurgery. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(5), 1263–1273 (2013)
11. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(7), 1409–1422 (2012)
12. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-backward error: Automatic detection of tracking failures. In: 20th International Conference on Pattern Recognition (ICPR), pp. 2756–2759 (2010)
13. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision 56(3), 221–255 (2004)
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. I-511–I-518 (2001)
15. Ozuysal, M., Fua, P., Lepetit, V.: Fast keypoint recognition in ten lines of code. In: IEEE Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2007)
16. Pickering, M.R., Muhit, A.A., Scarvell, J.M., Smith, P.N.: A new multi-modal similarity measure for fast gradient-based 2d-3d image registration. In: IEEE Engineering in Medicine and Biology Society, EMBC 2009, pp. 5821–5824 (2009)
17. Benhimane, S., Malis, E.: Homography-based 2d visual tracking and servoing. The International Journal of Robotics Research 26(7), 661–676 (2007)