Online Tracking and Retargeting with Applications to Optical Biopsy in Gastrointestinal Endoscopic Examinations

Menglong Ye, Stamatia Giannarou, Alexander Meining, Guang-Zhong Yang

PII: S1361-8415(15)00144-9
DOI: 10.1016/j.media.2015.10.003
Reference: MEDIMA 1046

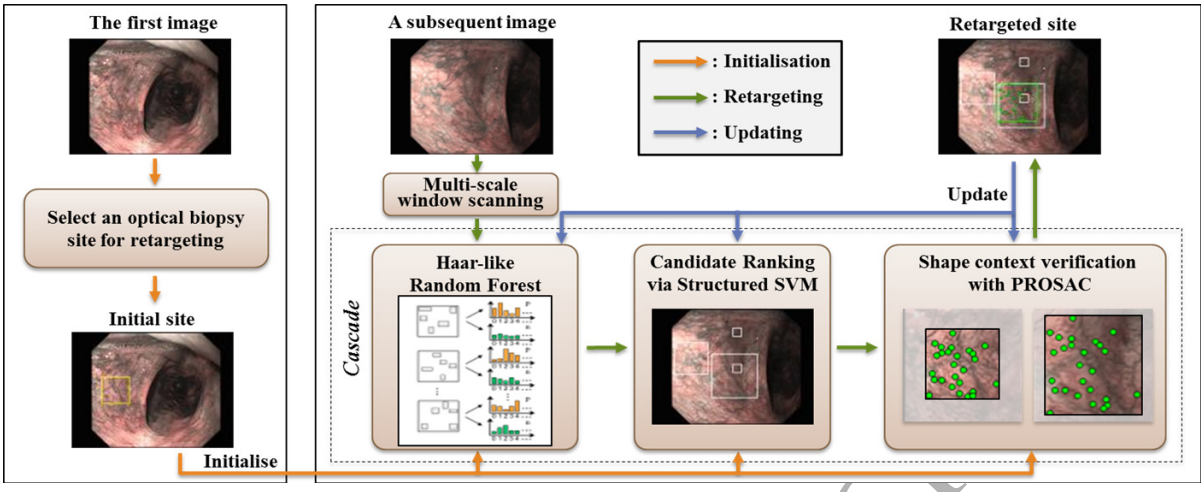
To appear in: *Medical Image Analysis*

Received date: 28 January 2015
Revised date: 30 September 2015
Accepted date: 2 October 2015

Please cite this article as: Menglong Ye, Stamatia Giannarou, Alexander Meining, Guang-Zhong Yang, Online Tracking and Retargeting with Applications to Optical Biopsy in Gastrointestinal Endoscopic Examinations, *Medical Image Analysis* (2015), doi: 10.1016/j.media.2015.10.003

Highlights

- An online detection cascade is introduced to address optical biopsy retargeting
- A random binary descriptor is proposed and used as a simple random forest classifier
- Shape context is combined with RANSAC to provide location verification for detection
- Detailed in-vivo validation showed that our framework outperforms existing trackers



Online Tracking and Retargeting with Applications to Optical Biopsy in Gastrointestinal Endoscopic Examinations

Menglong Ye[a,*], Stamatia Giannarou[a], Alexander Meining[b], Guang-Zhong Yang[a]

[a]*The Hamlyn Centre for Robotic Surgery, Imperial College London, United Kingdom*
[b]*Centre of Internal Medicine, Ulm University, Germany*

Abstract

With recent advances in biophotonics, techniques such as narrow band imaging, confocal laser endomicroscopy, fluorescence spectroscopy, and optical coherence tomography, can be combined with normal white-light endoscopes to provide *in vivo* microscopic tissue characterisation, potentially avoiding the need for offline histological analysis. Despite the advantages of these techniques to provide online optical biopsy *in situ*, it is challenging for gastroenterologists to retarget the optical biopsy sites during endoscopic examinations. This is because optical biopsy does not leave any mark on the tissue. Furthermore, typical endoscopic cameras only have a limited field-of-view and the biopsy sites often enter or exit the camera view as the endoscope moves. In this paper, a framework for online tracking and retargeting is proposed based on the concept of tracking-by-detection. An online detection cascade is proposed where a random binary descriptor using Haar-like features is included as a random forest classifier. For robust retargeting, we have also proposed a RANSAC-based location verification component that incorporates shape context. The proposed detection cascade can be readily integrated with other temporal trackers. Detailed performance evaluation on *in vivo* gastrointestinal video sequences demonstrates the performance advantage of the proposed method over the current state-of-the-art.

Keywords:
Tracking-by-detection, Tissue tracking, Structured SVM, Random binary descriptor, Optical biopsy retargeting, Gastrointestinal endoscopy

1. Introduction

Endoscopy is the standard technique for examining both the upper and lower gastrointestinal (GI) tracts. For upper GI endoscopy, it is mainly used for assessing abnormalities in the esophagus, stomach and duodenum. One of the common diseases in the upper GI tract is Barrett's esophagus, which is caused by chronic gastroesophageal reflux. Barrett's esophagus is widely associated with esophageal adenocarcinoma. For lower GI endoscopy, known as colonoscopy, the examination is mainly performed in the large intestine to investigate suspicious pathological sites, such as polyps, which may lead to colorectal cancer.

During endoscopy, tissue biopsies are frequently taken to provide definite pathological diagnosis of the target site. Due to practical constrains on tissue handling, the biopsy is only limited to a few target sites. Recently, optical biopsy has emerged as a promising tool for real-time *in situ* tissue characterisation. The main advantage of optical biopsy is its ability to assess pathologies non-invasively *in vivo* and *in situ*, avoiding the need of time-consuming histological analysis. Thus far, many biophotonic techniques have been developed. These include microscopic imaging or spectroscopy techniques such as confocal laser endomicroscopy, fluorescence spectroscopy, and optical coherence tomography (Hughes and Yang, 2012). They can be embodied in a mini-probe that can be readily inserted through the instrument channel of a normal endoscope. Enhancements of the endoscopic imaging system, such as Narrow Band

*Corresponding author. Tel: +44 (0)20 7594 1499.
Email address: menglong.ye11@imperial.ac.uk (Menglong Ye)

Imaging (NBI), have also been used to improve the visibility of sub-mucosal features. Optical biopsy has also be combined with robot-assisted endoscopy, and pioneering studies have been conducted in Newton et al. (2011, 2012) showing that robotic endoscopes can provide controlled contact of imaging probes with the tissue for acquiring high quality endomicroscopic images.

Despite its established benefit in non-invasive tissue characterisation, retargeting of optical biopsy sites is practically difficult, even when tissue biopsy is taken using forceps immediately after optical biopsy. This is because optical biopsy does not leave any visible marks on the tissue surface, which is further hampered by the rapid movement of the endoscope. The relatively small field-of-view (FOV) of the endoscope, coupled with the paucity of distinctive anatomical features, makes retargeting of previously identified biopsy sites difficult even for experienced observers. Furthermore, tissue deformation due to patient movement, peristalsis or respiration is another challenge to deal with in optical biopsy retargeting. For these reasons, there is a great demand clinically to develop a robust vision based technique for consistent retargeting of previously visited biopsy sites in GI endoscopic examinations. A direct application of this work is the retargeting of optical biopsies captured using probe-based confocal laser endomicroscopy. It is worth noting that our framework can be combined with any approach that identifies suspicious regions, either manually selected by users or by an automatic recognition system.

2. Related work and contributions

Thus far, several computer vision based techniques have been proposed to address the retargeting of optical biopsies. For example, Atasoy et al. (2009) presented a region-matching approach based on Markov random fields. In their work, affine-covariant regions combined with geometric constraints are used to facilitate retargeting of optical biopsy sites. Epipolar geometry has been used in Allain et al. (2012) such that for a query image, the biopsy is found by intersecting epipolar lines projected from a set of images where the biopsy site location is known. It should be noted that both of the above methods require multiple images that contain the same biopsy site. To facilitate sequential retargeting, a manifold approach based on laplacian eigenmaps (Atasoy et al., 2012) was proposed.

In addition to these offline approaches, online methods such as simultaneous localisation and mapping has been adopted by Mountney et al. (2009) to reconstruct a 3D map of salient features and the optical biopsy site. By tracking and localising the camera position, the method uses 3D-to-2D reprojection to retarget the optical biopsy site by assuming a rigid environment. The technique was subsequently extended to incorporate large rhythmic motion (Mountney and Yang, 2010). A more recent method has been proposed by Ye et al. (2013), which treats retargeting as a tracking-by-detection problem. A biopsy site is found in the query image by tracking and detecting its surrounding regions. Compared to previous approaches, this method does not need to generate image sets beforehand, as the initialisation and updating can be performed on-the-fly by adopting a tracking-by-detection paradigm (Kalal et al., 2012), thus making it clinically relevant. Another major advantage of this method is its robustness to global tissue deformation, which is challenging for previous techniques.

In this work, optical biopsy retargeting is addressed as a region tracking problem. For object tracking, a myriad of methods have been proposed in general computer vision. In Grabner et al. (2006), the authors propose to use an online adaptive boosting algorithm for tracking. The framework is able to adapt to object changes during tracking by updating a combination of weak classifiers with a single positive sample from the current target location and negative samples from the surrounding background. To deal with the drawback that the single positive sample might be suboptimal, a variant approach has been presented by Babenko et al. (2011), which trains weak classifiers with bags of positive and negative samples. An incremental principal component analysis (PCA) method has been adopted in Ross et al. (2008) that ensembles the object appearance into a low-dimensional space. During tracking, this subspace representation can be updated online. Other subspace learning methods for tracking are based on compressive sensing (Zhang et al., 2012) and sparse representation models (Zhong et al., 2012; Jia et al., 2012; Bao et al., 2012). To deal with deformation, Oron et al. (2012) proposed to use superpixels (Ren and Malik, 2003) to represent the object model, which enables tracking to be performed using locally orderless matching.

Recently, tracking-learning-detection (TLD) has been introduced in Kalal et al. (2012), which combines an optical-flow tracker with a cascaded detector. This approach enables training of random ferns (Ozuysal et al., 2010) and a nearest-neighbour classifier online in a boostraping scheme. However, as the approach relies on template matching at the classification stage, it is vulnerable to false positives. To deal with this drawback, an extended version has been proposed by Dinh et al. (2011) to exploit the local contextual information. Another tracking method is Struck,

Table 1: Summary of all evaluated methods. LBP represents local binary patterns and 'combined' represents 'generative + discriminative'.

Tracker	Appearance type	Model	Learning method
IVT (Ross et al., 2008)	holistic image intensity	generative	incremental PCA
TLD (Kalal et al., 2012)	LBP and image patches	discriminative	online random forest
CXT (Dinh et al., 2011)	LBP and image patches	discriminative	online random forest
SCM (Zhong et al., 2012)	local intensity histograms	combined	sparse coding
CT (Zhang et al., 2012)	Haar-like features	discriminative	random projections
ASLA (Jia et al., 2012)	local intensity histograms	generative	sparse coding
L1APG (Bao et al., 2012)	holistic image intensity	generative	sparse coding
OAB (Grabner et al., 2006)	Haar-like features	discriminative	online boosted trees
MIL (Babenko et al., 2011)	Haar-like features	discriminative	online boosted trees
Struck (Hare et al., 2011)	Haar-like features	discriminative	structured SVM
PSR (Ye et al., 2013)	LBP and image patches	discriminative	online random forest
OTR (the proposed)	Haar-like binary descriptor	discriminative	online random forest + structured SVM

recently proposed by Hare et al. (2011), which treats tracking as a prediction function of object transformations from successive images. In surgical vision, detection and tracking have been topics of particular interest, which include both surgical instrument tracking or detection (Sznitman et al., 2012, 2013, 2014; Reiter et al., 2014), and tissue tracking (Mountney et al., 2007; Mountney and Yang, 2008; Mountney et al., 2010; Richa et al., 2011, 2012; Giannarou et al., 2013).

In this paper, we propose an Online Tracking and Retargeting (OTR) framework for optical biopsy. Our previous work on Pathological Site Retargeting (PSR) (Ye et al., 2013) has demonstrated that by tracking surrounding regions, the biopsy sites can be retargeted reliably. However, the requirement of tracking multiple regions can introduce a relatively high computational cost. Furthermore, the local planar assumption used in our previous work can result in low recall values for in vivo applications. The purpose of this work is to remove these constraints without compromising the performance of tracking and retargeting. Instead of tracking multiple regions, the OTR method draws information only from a single image area by adopting local shape context, thus relaxing the planar constraints to cater for free-form deformation. The low recall values in our previous method have been mitigated and validation has been conducted on *in vivo* GI videos compared to ten state-of-the-art trackers and our previous PSR method. A summary of the evaluated trackers is provided in Table 1.

The main contributions of this work are: 1) a random binary descriptor inspired by Haar-like features is proposed; 2) a new detection cascade is introduced, which incorporates an online random forest classifier, a structured Support Vector Machine (SVM) and a verification component based on Random Sampling Consensus (RANSAC); 3) for verification, shape context information and location refinement are combined with RANSAC, thus improving the robustness to false positives. The *in vivo* datasets along with the ground truth are made available online (http://hamlyn.doc.ic.ac.uk/vision) to facilitate the evaluation of existing and future tissue tracking techniques by the medical image computing community.

3. Methods

In this work, retargeting is formulated as a tracking-by-detection problem. An optical biopsy site is defined by the user, which is then tracked in the rest of the image sequence. To generate candidates of this site on subsequent images, multi-scale window scanning (Viola and Jones, 2001) is applied. More specifically, the image is scanned horizontally and vertically using bounding boxes with a predefined displacement step. To deal with the scale changes of the biopsy site, the bounding boxes are set to varying sizes during scanning, such that a set of image patches is generated. These image patches can be treated as the candidates of the biopsy site. However, further filtering is required to extract the best candidate that represents the site.

In this section, we focus on introducing a new detection cascade, which includes a classifier based on random binary descriptors, a structured SVM that ranks the candidate locations of the site, and a verification component that

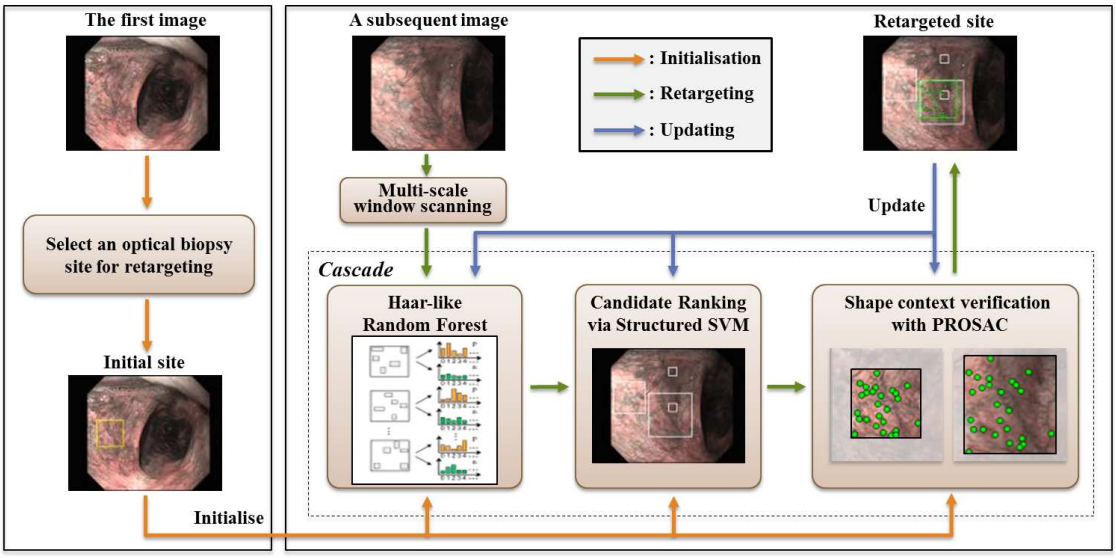


Figure 1: An overview of the proposed detection cascade. An optical biopsy site is selected in the first image to initialise the cascade so that retargeting can be performed in the subsequent images in a video sequence.

uses a variant of RANSAC to find the best image patch. An overview of the detection cascade has been provided in Fig. 1. Our detection cascade can be easily combined with any temporal tracking method such as (Kalal et al., 2010).

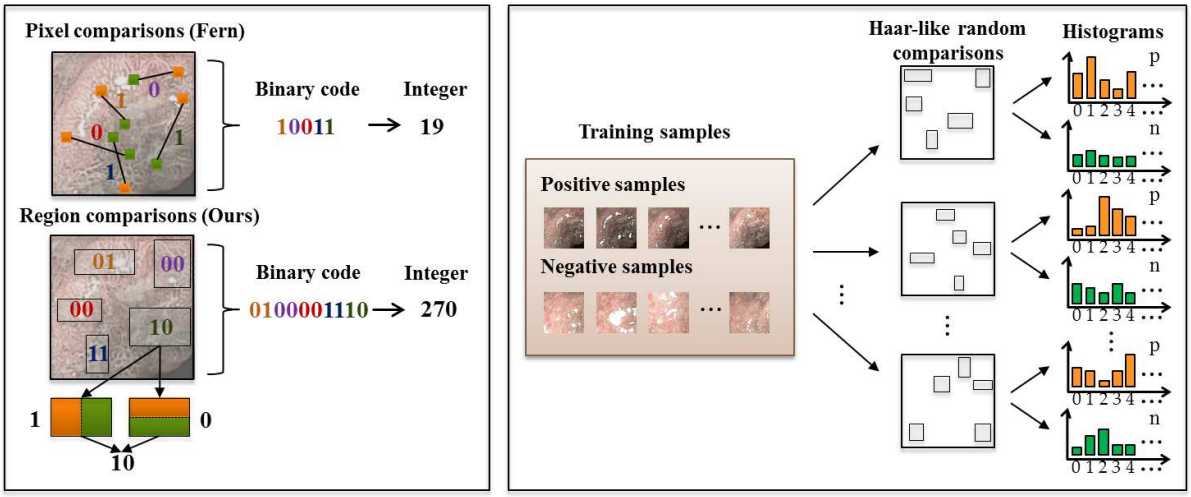
3.1. Haar-like random forest classifier

Recent uses of binary descriptors have enabled computationally efficient object tracking and detection (Calonder et al., 2012; Leutenegger et al., 2011; Rublee et al., 2011). This is because binary comparisons of local pixels are fast and robust to illumination changes. In this paper, we introduce a Haar-like random binary descriptor based on local region comparisons. It is worth noting that random ferns (Ozuysal et al., 2010) is also a random binary descriptor. The difference between our descriptor and the one used by Ozuysal et al. (2010) is that our method computes binary codes based on local region comparisons rather than pixel comparisons, which is more robust to image noise while maintaining the efficiency of binary descriptors at the same image scale.

3.1.1. Random binary descriptor

To extract the descriptor of an image patch, we consider multiple sets of rectangles inside the image patch. A rectangle configuration is denoted as $\{x, y, w, h\}$, where x and y are the centre position coordinates and w and h are the width and height. Each rectangle is then divided into two pairs of regions: the left and right halves, and the top and bottom halves. We calculate the sums of the pixel intensities using integral images in these four regions, which are denoted as S_l, S_r, S_t and S_b for the left, right, top and bottom regions, respectively. By comparing these sums, a two-digit binary code can be obtained. For instance, binary code 10 is obtained when $S_l \geq S_r$ and $S_t < S_b$. In this work, the values of the rectangle configuration are randomly generated. Therefore, our proposed method can be performed in a similar manner to random ferns but with improved robustness to pixel noise with the same image scale setting, as shown in Fig. 2a. An alternative approach to achieving region comparisons would be to downsize the image according to the proposed rectangle configurations, and then perform pixel comparisons. However, this would be computationally inefficient, because these configurations are randomly generated, requiring multiple resizing operations.

In a typical setting, we generate M sets of rectangles and each set has Z rectangles. Once generated, the rectangle configuration remains fixed during retargeting. A binary code d_m of $2Z$ digits can be obtained from set m. Finally, the image patch can be described with a binary code $\mathbf{d} = [d_1, d_2, \ldots, d_M]$ of $2MZ$ digits by concatenating the binary codes from the M rectangle sets.



(a) Difference between random ferns and the proposed random binary descriptor.

(b) Training Haar-like binary descriptor as a simplified random forest classifier.

Figure 2: Illustration of the Haar-like random forest classifier.

In this work, $\mathbf{d}$ is used in two ways. The first is to convert each d_m into an integer number, which can then be used to formulate a simplified random forest classifier. The second is to combine the descriptor with a structured SVM for candidate ranking, which will be introduced in Section 3.2.

3.1.2. A simplified random forest classifier

To formulate a simplified random forest classifier, we convert every binary code into an integer number. Given the positive and negative samples of the tracked site on the first image of the endoscopic sequence, we generate for each of the M rectangle sets a pair of histograms ranging from 0 to $2^{2Z} - 1$ (see Fig. 2b). These histograms represent the distribution of the integer numbers in the positive and negative samples for each set of rectangles. When analysing a query image patch $\mathbf{y}_j$, the M rectangle sets are applied to generate the binary codes and convert them to integer numbers. To evaluate the probability of $\mathbf{y}_j$ being a true representation of the optical biopsy site, the posterior probabilities for every d_m are estimated as $P(\mathbf{y}_j|d_m) = \frac{p}{p+n}$, where p and n are the frequencies of the integer corresponding to the code d_m in the positive and negative histograms, respectively. The final probability is evaluated as:

$$P(\mathbf{y}_j|\mathbf{d}) = \frac{1}{M}\sum_{m=1}^{M} P(\mathbf{y}_j|d_m). \tag{1}$$

Note that when $p = n = 0$, we assign $P(\mathbf{y}_j|d_m) = 0$ to avoid the denominator being zero. By setting a threshold θ, Eq. 1 can then be used to filter out the image patches with low confidence values. It should be noted that when combining the base classifiers, the independence of these classifiers should be guaranteed (Breiman, 2001). Therefore, in practice, when performing random generation of sets of rectangles, we control the sizes of the rectangles to ensure they do not overlap with each other. As described earlier, we use the proposed random binary descriptors to construct a simplified random forest classifier which is referred to as Haar-like Random Forest (HRF) in the remainder of this paper. This represents the first component in the aforementioned cascade filtering scheme.

3.2. Ranking candidates using structured SVM

In this work, multi-scale window scanning is applied to generate the initial image patches, and HRF is used to retain a set of candidate patches $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L\}$, where $\mathbf{y}_l = \left(r_l^x, r_l^y, r_l^w, r_l^h\right)$ describes the centre position and size of the patch in an image. To refine the set $\mathcal{Y}$, we adopt a structured SVM, to rank the candidates, and retain the top ones which will be used for further processing.

3.2.1. Structured SVM formulation

Structured SVMs have recently been adopted in computer vision for object detection and tracking (Blaschko and Lampert, 2008; Bertelli et al., 2011; Hare et al., 2011; Yao et al., 2012). In this work, our aim is to learn a function $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, such that given an image $\mathbf{x} \in \mathcal{X}$, f can be used to estimate the best patch to represent the optical biopsy site. This is defined as:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}). \tag{2}$$

where $\mathcal{X}$ represents all the images in the video sequence and $\mathcal{Y}$ represents all the candidate patches in image $\mathbf{x}$. $f(\mathbf{x}, \mathbf{y})$ is considered as a linear function:

$$f(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}) \rangle , \tag{3}$$

where $\langle \cdot, \cdot \rangle$ represents the dot product operation, and $\Phi(\mathbf{x}, \mathbf{y})$ is the feature representation of patch $\mathbf{y}$ in image $\mathbf{x}$. Here, we assign the descriptor $\mathbf{d}$ (Section 3.1) to $\Phi(\mathbf{x}, \mathbf{y})$. Therefore, learning $f(\mathbf{x}, \mathbf{y})$ can be achieved by learning the weight vector $\mathbf{w}$. When T images are available, the learning process can be approached as an optimisation problem:

$$\begin{aligned} &\min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N_t}\sum_{i=1}^{N_t}\xi_{t,i}\right) \\ &s.t. \forall t, i : \xi_{t,i} \geq 0 \\ &\quad \forall t, i : \left\langle \mathbf{w}, \Phi\left(\mathbf{x}_t, \mathbf{y}_t^{\star}\right)\right\rangle - \left\langle \mathbf{w}, \Phi\left(\mathbf{x}_t, \mathbf{y}_{t,i}\right)\right\rangle \geq \Delta\left(\mathbf{y}_t^{\star}, \mathbf{y}_{t,i}\right) - \xi_{t,i}. \end{aligned} \tag{4}$$

The aim of the equation above is to optimise $\mathbf{w}$, such that the margin between the best solution and any other solutions can be maximised. Here, we denote $\mathbf{x}_t$ as the image at time t and $\mathbf{y}_{t,i}$ as the image patch i in $\mathbf{x}_t$ where $i = 1 \ldots N_t$. $\Delta(\cdot, \cdot)$ is a problem-specific loss function, λ is a constant scaling parameter, and $\xi_{t,i}$ is a slack variable. $\mathbf{y}_{t,i} \neq \mathbf{y}_t^{\star}$ and $\mathbf{y}_t^{\star}$ is the best patch that represents the biopsy site at time t, which in our case, $\mathbf{y}_t^{\star}$ has been verified by shape context as explained in Section 3.3.

In our case, effective online learning is required, therefore we adopt the mini-batch Pegasos algorithm (Shalev-Shwartz et al., 2011), which is a stochastic gradient descent method for structured prediction. We convert Eq. 4 into an unconstrained form at time t:

$$\begin{aligned} \min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{N_t}\sum_{i=1}^{N_t} \ell\Big(\Delta\left(\mathbf{y}_t^{\star}, \mathbf{y}_{t,i}\right) \\ + \langle \mathbf{w}, \Phi\left(\mathbf{x}_t, \mathbf{y}_{t,i}\right)\rangle - \left\langle \mathbf{w}, \Phi\left(\mathbf{x}_t, \mathbf{y}_t^{\star}\right)\right\rangle\Big), \end{aligned} \tag{5}$$

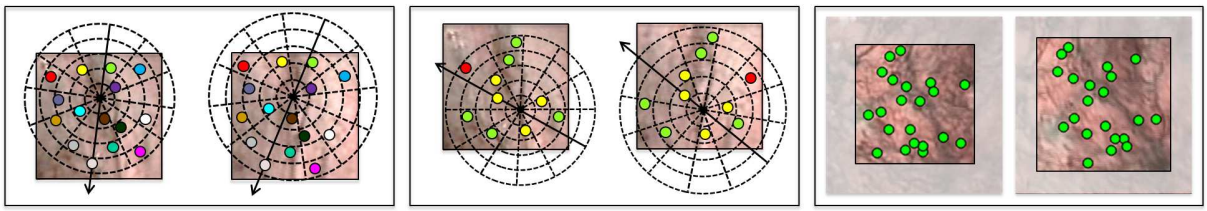
where $\ell(a) = \max(0, a)$. An objective can be then defined as

$$g(\mathbf{x}, t) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{N_t}\sum_{i=1}^{N_t} \ell\left(\Delta\left(\mathbf{y}_t^{\star}, \mathbf{y}_{t,i}\right) - \langle \mathbf{w}, \delta\Phi_t(\mathbf{y})\rangle\right), \tag{6}$$

with $\delta\Phi_t(\mathbf{y}) = \Phi\left(\mathbf{x}_t, \mathbf{y}_t^{\star}\right) - \Phi\left(\mathbf{x}_t, \mathbf{y}_{t,i}\right)$. To estimate $\mathbf{w}_{t+1}$, the sub-gradient is derived as

$$\nabla_t = \lambda \mathbf{w}_t - \frac{1}{N_t}\sum_{i=1}^{N_t} \mathbb{1}\left(\Delta\left(\mathbf{y}_t^{\star}, \mathbf{y}_{t,i}\right) > \langle \mathbf{w}_t, \delta\Phi_t(\mathbf{y})\rangle\right)\delta\Phi_t(\mathbf{y}). \tag{7}$$

Here $\mathbb{1}(\cdot)$ is an indicator function. To update $\mathbf{w}$, we perform the gradient descent step as $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla_t$, where $\eta_t = \frac{1}{\lambda t}$ is the step size. At time $t + 1$, the weight $\mathbf{w}_{t+1}$ is used and $\mathbf{w}_{t+2}$ is estimated at the end of the time step.

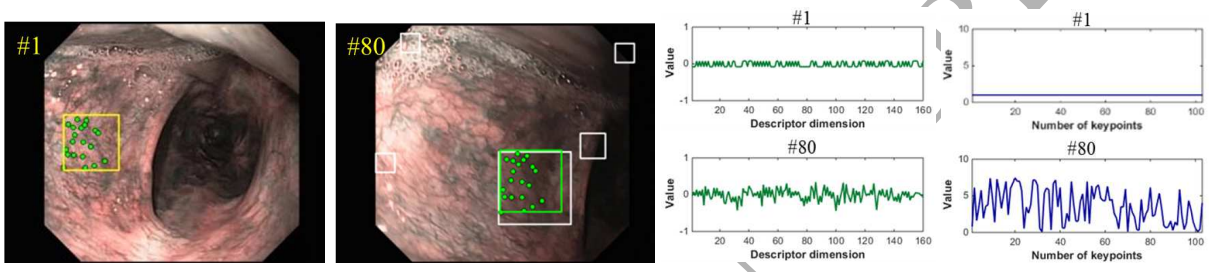


(a) If the keypoints are perfectly matched (in same color), they should lie in the same zones in the polar grids.

(b) To apply the RANSAC algorithm, a subset of correspondences (yellow) is sampled to identify the inliers (green) and outliers (red).

(c) An example result of location verification using shape context with PROSAC. The inliers of correspondences are shown in green.

Figure 3: Illustration of the combination of shape context with RANSAC.



(a) The image where an optical biopsy site is selected.

(b) The image where the site is re-targeted from candidates and refined using keypoint matches.

(c) Weight vector $\mathbf{w}$ (binary descriptor is normalised) at two timestamps.

(d) Weight vector $\mathbf{v}$ (data points are norms of descriptors) at two timestamps.

Figure 4: Examples of location refinement and weight vectors $\mathbf{w}$ and $\mathbf{v}$ at two timestamps.

3.2.2. Ranking candidate patches

In the proposed framework, Eq. 2 is used to select a set of top candidate patches from $\mathcal{Y}$ for every image. This is different from the methods proposed in Hare et al. (2011); Yao et al. (2012) where only the top patch is chosen. It is important to note that these methods perform structured prediction by sampling image patches near the previous tracked object location. This sampling strategy is reasonable when the object consistently appears in subsequent images, however it is not reliable when the object re-enters the FOV after disappearance. Therefore, we propose to search for the object in the entire image using HRF. However, when searching in the entire image domain, the top ranked patch from the aforementioned structured prediction scheme is not always the best patch to describe the object (the top ranked patch might be suboptimal). This is because structured SVM requires a certain number of iterations to converge to the batch SVM when being trained online, as explained in Shalev-Shwartz et al. (2011).

In this regard, we select the top K candidates from $\mathcal{Y}$ to cater for convergence issues. Eq. 3 is used to rank patches in $\mathcal{Y}$, and retain the top ones in $\hat{\mathcal{Y}} = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_K\}$. In order to find the best estimate $\mathbf{y}^\star$ from all the candidates, at the next stage, a variant of the RANSAC algorithm called Progressive Sampling Consensus (PROSAC) (Chum and Matas, 2005) is used to verify the local shape context described by visual features.

3.3. Location verification using shape context

In our previous work (Ye et al., 2013), we assume that the local area on the tissue surface is planar. Therefore, a homography was estimated to transform the corresponding points inside the area from successive images. However, this assumption is vulnerable for areas with high surface curvatures. To circumvent this problem, we choose to combine the PROSAC algorithm with local shape context (Belongie et al., 2002) using fast-Hessian features (Bay et al., 2006).

3.3.1. Shape context with PROSAC

Shape context is a well-established method for object recognition based on shape information (Belongie et al., 2002). In this work, the shape context of an image is described using keypoints, which is then further combined with PROSAC for robust verification. Given a set of keypoints detected in an image patch, we sample a subset of keypoints and use them to create a polar grid to describe the spatial distribution of all the keypoints in the patch. The grid is composed of 24 equally spaced angular bins and radial bins of 10-pixel radius each. The origin of the grid is the centre of mass of the keypoint subset, which is found by averaging the coordinates of the keypoints in the subset. The reference orientation of the grid can be obtained by connecting the centre of mass with any keypoint in the subset. The relative scale of the grid is estimated using the distances between the centre of mass to the keypoint that was selected to define the reference orientation.

Given another image patch for comparison, the two patches are compared in a RANSAC-like algorithm, which is divided into two stages, namely preparation and iteration stages, respectively. During the preparation stage, keypoints are detected in both image patches and described using the proposed random binary descriptor $\mathbf{d}$. The matching scores (dot products) between the keypoint descriptors are calculated, and the keypoint correspondences with high scores are retained. In the iteration stage, a subset of keypoint correspondences is randomly selected to create two polar grids for the two patches. The reference orientation and scale for the two polar grids are estimated using corresponding keypoints. Then, all the keypoints correspondences of the patches are examined and a correspondence is marked as an inlier if the keypoints are located in the same zone in the polar grids. If the number of inliers is larger than a predefined threshold, the patches are identified as the same. Otherwise, the above process is iterated until a maximum number of iterations has been reached.

To use the above mentioned approach for biopsy location verification, Hessian features (Bay et al., 2006) are detected inside the optical biopsy site that has been selected for retargeting. These features are treated as model keypoints $\{u_i\}_{i=1}^{I}$. For a subsequent image $\mathbf{x}_t$, we obtain $\hat{\mathcal{Y}}_t$ as described in Section 3.2. For a given patch in $\hat{\mathcal{Y}}_t$, keypoints are detected and matched with the model keypoints. If the compared patches match, their shape contexts should be similar and the number of inliers should be larger than a predefined threshold. These details are illustrated in Figs. 3a and 3b.

The PROSAC algorithm is a variant of RANSAC that sorts all the keypoint correspondences in a descending order according to their matching scores before sampling. The advantage of PROSAC compared to traditional RANSAC is that the former encourages the selection of high score correspondences during sampling after sorting the scores. This enables the method to reach the optimal solution with fewer iterations (Chum and Matas, 2005). For each sample set, inliers and outliers are identified. The best set of correspondences is the sample set with the largest number of inliers. An example result of the proposed PROSAC with shape context has been provided in Fig. 3c, where inlier correspondences are shown in green.

3.3.2. Accommodating for keypoint appearance changes

If we keep the descriptors of the model keypoints static, these descriptors would not be robust enough to handle dynamic changing environment of the endoscopic video. Thus, we update the model keypoint descriptors on-the-fly via the structured SVM formulation described earlier. A similar work that uses the structured SVM for descriptor updating has also been presented by Hare et al. (2012), and the experiments have shown that it outperforms the static keypoint model. At this stage, we adopt the basic Pegasos algorithm instead of mini-batch Pegasos, due to its computational efficiency. The objective for online descriptor updating is defined as

$$g(\mathbf{v}, t) = \frac{\lambda}{2}\|\mathbf{v}\|^2 + \ell\left(\Delta\left(\mathbf{h}_t^{\star}, \mathbf{h}_t'\right) - \langle \mathbf{v}, \delta\Psi_t(\mathbf{h})\rangle\right) + \sum_{(u_i, v_j)\in \mathbf{c}_t^{\star}} \ell\left(1 - \left\langle \mathbf{v}_i, \mathbf{d}_j^{\star} - \mathbf{d}_j'\right\rangle\right). \tag{8}$$

Here, $\delta\Psi_t(\mathbf{h}) = \langle \mathbf{v}, \Psi(\mathbf{c}_t, \mathbf{h}_t^{\star})\rangle - \langle \mathbf{v}, \Psi(\mathbf{c}_t, \mathbf{h}_t')\rangle$, where $\mathbf{h}$ is the transform between the original and current centres of the biopsy site. $\mathbf{h}_t^{\star}$ and $\mathbf{h}_t'$ are the best and second best transforms estimated from PROSAC at time t, respectively. The second best transform is obtained from the correspondence set with the second highest number of inliers. $\mathbf{c}_t$ represents a set of correspondences $\left(u_i, v_j\right)$ between the model keypoints and the keypoints at time t, and $\mathbf{c}_t^{\star}$ is the best correspondence set from PROSAC. In this section, $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_I]$ is a weight vector that concatenates all the

Algorithm 1 Cascade Retargeting

Input: $\mathbf{x}_t$, $\mathbf{w}_t$, $\mathbf{v}_t$, T_c
Output: $\mathbf{y}_t^\star$, $\mathbf{w}_{t+1}$, $\mathbf{v}_{t+1}$

if Initialisation is needed **then**
  Retrieve samples in relation to the selected biopsy site to train the HRF classifier
  Assign the random binary descriptor of the site to $\mathbf{w}_0$
  Assign the random binary descriptors of the keypoints inside the site to $\mathbf{v}_0$
else
  Generate initial image patches from image $\mathbf{x}_t$ using multi-scale window scanning;
  Obtain candidate patches $\mathcal{Y}_t$ with the HRF (Eq. 1);
  Estimate top patches $\hat{\mathcal{Y}}_t = \{\hat{\mathbf{y}}_{t,1}, \hat{\mathbf{y}}_{t,2}, \ldots, \hat{\mathbf{y}}_{t,K}\}$ with $\mathbf{w}_t$ using Eq. 3;
  Set $\mathbf{c}_t^\star = \varnothing$
  for $k = 1, 2, \ldots, K$ **do**
    Extract keypoints in $\hat{\mathbf{y}}_{t,k}$;
    Obtain $\mathbf{c}_{t,k}$ using $\mathbf{v}_t$ via PROSAC combined with shape context;
    if $|\mathbf{c}_{t,k}| > T_c$ **then**
      Set $\mathbf{c}_t^\star = \mathbf{c}_{t,k}$
      break;
    end if
  end for
  if $\mathbf{c}_t^\star = \varnothing$ **then**
    Set $\mathbf{y}_t^\star = \varnothing$
    Perform gradient descent to generate $\mathbf{w}_{t+1}$ and $\mathbf{v}_{t+1}$;
  else
    Estimate $\mathbf{h}_t^\star$ using $\mathbf{c}_t^\star$
    Obtain $\mathbf{y}_t^\star$ via location refinement;
    Perform gradient descent to generate $\mathbf{w}_{t+1}$ and $\mathbf{v}_{t+1}$;
    Update the HRF classifier;
  end if
end if

model keypoint descriptors. The random binary descriptor described in Section 3.1 is used here for the representation $\Psi(\mathbf{c}, \mathbf{h})$ of the keypoints.

The objective of basic Pegasos in Eq. 8 is an approximation, as it only considers training using the best and second best samples while the mini-batch Pegasos algorithm requires a set of training samples for each iteration. This is particularly efficient when it is time-consuming to obtain mini-batch training samples (Shalev-Shwartz et al., 2011), e.g., in a RANSAC-like algorithm. Another important feature in Eq. 8 is that it encourages inlier correspondences by increasing the distance between the best descriptor $\mathbf{d}_j^\star$ and the second best $\mathbf{d}_j'$, as mentioned in Hare et al. (2012).

The weight vector $\mathbf{v}$ is updated via $\mathbf{v}_{t+1,i} = \mathbf{v}_{t,i} - \eta_t \nabla_t$, where the sub-gradient is obtained as follows:

$$\begin{aligned}\nabla_t = \lambda \mathbf{v}_{t,i} - \mathbb{1}\left(\Delta\left(\mathbf{h}_t^\star, \mathbf{h}_t'\right) > \left\langle \mathbf{v}_{t,i}, \delta\Psi_t(\mathbf{h})\right\rangle\right)\delta\Psi_t(\mathbf{h}) \\ - \mathbb{1}\left(\left(u_i, v_j\right) \in \mathbf{c}_t^\star\right)\mathbb{1}\left(1 > \left\langle \mathbf{v}_{t,i}, \mathbf{d}_j^\star - \mathbf{d}_j'\right\rangle\right)\left(\mathbf{d}_j^\star - \mathbf{d}_j'\right).\end{aligned} \tag{9}$$

When processing image $\mathbf{x}_t$, the matching scores $\left\langle \mathbf{v}_{t,i}, \mathbf{d}_j \right\rangle$ are calculated between the model keypoints and the keypoints inside the image patch $\hat{\mathbf{y}}_{t,k} \in \hat{\mathcal{Y}}_t$. These matching scores are then used for PROSAC to reach an optimal solution $\mathbf{h}_t^\star$.

It is worth noting that in this framework, we have used Generalised Hough Transforms (GHTs) to represent $\mathbf{h}$, $\mathbf{h}_t'$ and $\mathbf{h}_t^\star$. Compared to using homographies (Ye et al., 2013) that are error-prone to non-planar surface, GHTs do not rely on the planar assumption and provide more freedom to cater for tissue curvatures. Therefore, once the number of inliers is greater than a predefined threshold T_c, we use the inliers to generate the transform $\mathbf{h}_t^\star$ via a spatial voting

scheme using the keypoints similar to that proposed by Nebehay and Pflugfelder (2014). After $\mathbf{h}_t^\star$ is obtained, with its associated correspondence set $\mathbf{c}_t^\star$, a gradient descent step can be performed to obtain $\mathbf{v}_{t+1}$. In addition to this, $\mathbf{h}_t^\star$ and $\mathbf{c}_t^\star$ are further used to refine the biopsy site location on $\mathbf{x}_t$.

3.3.3. Location refinement

To further refine the target location, we firstly compute the centre position $\left(r_t^x, r_t^y\right)$ by transforming the original position $\left(r_0^x, r_0^y\right)$ with $\mathbf{h}_t^\star$. To calculate the width r_t^w and height r_t^h, we consider the changes of the pairwise keypoint distances from time instant 0 to t. Given two arbitrary pairs of inliers $\left(u_i, v_j\right)$ and $\left(u_{i'}, v_{j'}\right)$ in $\mathbf{c}_t^\star$, the euclidean distances $\|u_i - u_{i'}\|$ and $\|v_j - v_{j'}\|$ are calculated. A scale factor can be then computed as

$$s = \frac{\|v_j - v_{j'}\|}{\|u_i - u_{i'}\|} \tag{10}$$

After computing all the pairwise distances in $\mathbf{c}_t^\star$, a set of scale factors $\mathcal{S}$ can be obtained. The final scale factor is chosen as the median of this set $s^\star = median\,(\mathcal{S})$. Therefore, the width and height are assigned as $r_t^w = s^\star r_0^w$ and $r_t^h = s^\star r_0^h$, where w_0 and h_0 are the original width and height, respectively. With this, the refined site patch $\mathbf{y}_t^\star = \left(r_t^x, r_t^y, r_t^w, r_t^h\right)$ is now obtained (the optical biopsy site is retargeted).

The refined site patch is then used to update $\mathbf{w}$ with the sub-gradient in Eq. 7, as well as to generate positive and negative image patches to online train the HRF classifier. Examples of location refinement and weight vectors $\mathbf{w}$ and $\mathbf{v}$ have been provided in Fig. 4. The overall procedure of our proposed cascade scheme is summarised in Algorithm 1.

3.4. Combination with a temporal tracker

The proposed online detection cascade can be readily combined with any temporal object trackers. In this paper, we show a case of combining it with the forward-backward (FB) tracker (Kalal et al., 2010). It is worth noting that the FB tracker has an error detection component embedded, based on analysing the FB errors of optical flow. And this helps stopping tracking when tracks are not reliable. At time t, the FB tracker can provide an estimate of the site location given it is visible at time $t-1$. Since our detection cascade provides accurate estimation of the biopsy site location using the aforementioned shape context information, whenever there is an estimate $\mathbf{y}_t^\star$ from Algorithm 1, it can be treated as the optimal location to correct the FB tracker to track the site from time t to $t+1$. When the detection cascade provides an estimate, this estimate will be used as the retargeting result, and the result from tracking is then ignored. When the detection cascade does not give an estimate, the result from tracking will be used as the retargeting result. This simple combination is different from TLD that fuses tracking and detection by averaging their results. Experimental results have verified that our combination results in lower location errors.

4. Experimental Results

In this section, we present the performance evaluation of the proposed OTR approach for optical biopsy retargeting. For comparison, 10 state-of-the-art trackers (summarised in Table 1) with publically available implementations have been used. We have also compared OTR with our previously published PSR method (Ye et al., 2013). As PSR only estimates the centres of the biopsy sites, the comparison was performed using the distance between the estimated and the ground truth centre locations. All the algorithms have been installed on a HP Z800 workstation (with 24GB RAM, Intel Xeon x5650 CPU and Nvidia GeForce GTX 770). The *in vivo* videos used for evaluation were collected by using Olympus NBI and Pentax i-scan endoscopes. Several components in our framework have been parallelised using CUDA GPU programming, which include the scanning-windows, the HRF classifier and the PROSAC shape context. The proposed method currently can achieve average 23 frames-per-second for videos with image size of 640x480 pixels (see Table 2).

Table 2: Average frames-per-second (FPS) of the proposed framework performed on 10 *in vivo* GI sequences.

Seq.	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	**Overall**
Images	705	1003	1700	336	493	325	1349	578	266	1139	7894
Average FPS	20.25	21.80	22.08	24.00	23.48	23.21	20.13	26.27	24.18	26.49	22.78

4.1. Configuration for initialisation and online training

For the HRF classifier, we generate $M = 8$ sets of rectangles and each individual set contains $Z = 10$ rectangles. To initialise the proposed HRF classifier, the initial positive and negative training sets are generated from affine warps of the selected biopsy site and background regions, respectively (Kalal et al., 2012; Dinh et al., 2011; Ye et al., 2013). It should noted that the affine-warped positive samples enable the robustness of our framework to in-plane rotation changes of a biopsy site. For updating HRF online, the training sets are generated when $\mathbf{y}_t^\star \neq \varnothing$ in Algorithm 1. The online positive samples are the affine warps of $\mathbf{y}_t^\star$, and the negative samples are chosen from $\hat{\mathcal{Y}}$ that have overlap smaller than 0.5 to $\mathbf{y}_t^\star$. The overlap is the metric used in PASCAL VOC challenge (Everingham et al., 2010) and defined as

$$o\left(\mathbf{y}_1, \mathbf{y}_2\right) = \frac{\mathbf{y}_1 \cap \mathbf{y}_2}{\mathbf{y}_1 \cup \mathbf{y}_2}, \tag{11}$$

which represents the overlap ratio between rectangular regions $\mathbf{y}_1$ and $\mathbf{y}_2$. Here, symbols '$\cap$' and '$\cup$' are the intersection and union operations on the rectangular regions.

The initial binary descriptors of the selected biopsy site and its model keypoints are used to initialise $\mathbf{w}$ and $\mathbf{v}$. For updating $\mathbf{w}_t$ and $\mathbf{v}_t$ online, the loss functions $\Delta\left(\mathbf{y}_t^\star, \mathbf{y}_{t,i}\right)$ and $\Delta\left(\mathbf{h}_t^\star, \mathbf{h}_t'\right)$ need to be specified. Here, we define $\Delta\left(\mathbf{y}_t^\star, \mathbf{y}_{t,i}\right) = 1 - o\left(\mathbf{y}_t^\star, \mathbf{y}_{t,i}\right)$, and $\Delta\left(\mathbf{h}_t^\star, \mathbf{h}_t'\right) = \|\mathbf{c}_t^\star| - |\mathbf{c}_t'\|$ where $|\mathbf{c}_t^\star|$ and $|\mathbf{c}_t'|$ are the numbers of inliers for $\mathbf{h}_t^\star$ and $\mathbf{h}_t'$, respectively. We retain the top $K = 10$ candidate patches for every image after ranking (Section 3.2), and set $\lambda = 0.1$ for both Eq. 7 and Eq. 9. When applying shape context with PROSAC for location verification, $|\mathbf{c}_t^\star|$ has to be greater than $T_c = I/6$, which is an arbitrary threshold, determined given the initial number of keypoints I inside the selected biopsy site.

4.2. Metrics for evaluation

In this work, two evaluation metrics are used, namely, the overlap ratio (Eq. 11) and the centre location distance between the ground truth and the tracking results. The quantitative comparisons of our method to the other trackers have been performed on 10 *in vivo* GI sequences (7894 images in total). The initial locations of the optical biopsy sites chosen to be tracked correspond to either pathological sites (small polyps) or regions with complex vascular information, which are difficult for clinicians to remember. The initial patch centres and sizes for tracking were defined by a GI expert. Inspired by the benchmarking work in Wu et al. (2013), we have manually labelled the ground truth, and presented the results in a similar manner. The performance of all the trackers were assessed using the average centre location error, precision and recall values, as well as the F-measure. For a sequence i processed by a tracker j, the precision value $\alpha_{i,j}$ is calculated as the number of true positives divided by the number of tracking results:

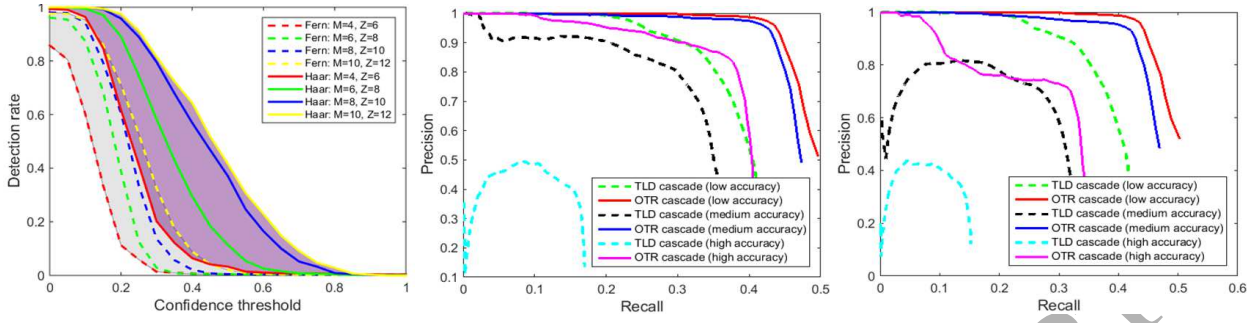
$$\alpha_{i,j} = \frac{|\mathcal{D}_{i,j} \cap \mathcal{G}_i|}{|\mathcal{D}_{i,j}|}, \tag{12}$$

where $\mathcal{D}_{i,j}$ is the set of tracking results, and $\mathcal{G}_i$ is the set of the ground truth. Here, the symbol '$\cap$' represents the intersection operation between the two sets. The recall $\beta_{i,j}$ is then defined as

$$\beta_{i,j} = \frac{|\mathcal{D}_{i,j} \cap \mathcal{G}_i|}{|\mathcal{G}_i|}, \tag{13}$$

which is the number of true positives divided by the number of images where the biopsy site is visible. And the F-measure $\gamma_{i,j}$ is the harmonic mean of precision and recall:

$$\gamma_{i,j} = \frac{2 \cdot \alpha_{i,j} \cdot \beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}}. \tag{14}$$



(a) Comparison between the proposed HRF and random ferns in terms of M, Z and θ parameter settings. Our proposed HRF consistently outperforms random ferns.

(b) Precision-recall curves of detectors using overlap.

(c) Precision-recall curves of detectors using location error.

Figure 5: Performance evaluation of the proposed detection cascade.

Table 3: Quantitative results of our online detection cascade compared to the TLD cascade. **Bold** numbers represent better performance.

Accuracy scenarios	TLD cascade			**OTR cascade**		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Low	0.92	0.29	0.44	**0.99**	**0.41**	**0.58**
Medium	0.84	0.26	0.40	**0.97**	**0.40**	**0.57**
High	0.43	0.13	0.20	**0.87**	**0.36**	**0.51**

It is worth noting that since two metrics are used in this study, the true positives can be determined either by the overlap ratio or the centre location error. That is, a tracking result can be marked as true positive when its overlap ratio with the ground truth is larger or the centre location distance is smaller than predefined thresholds.

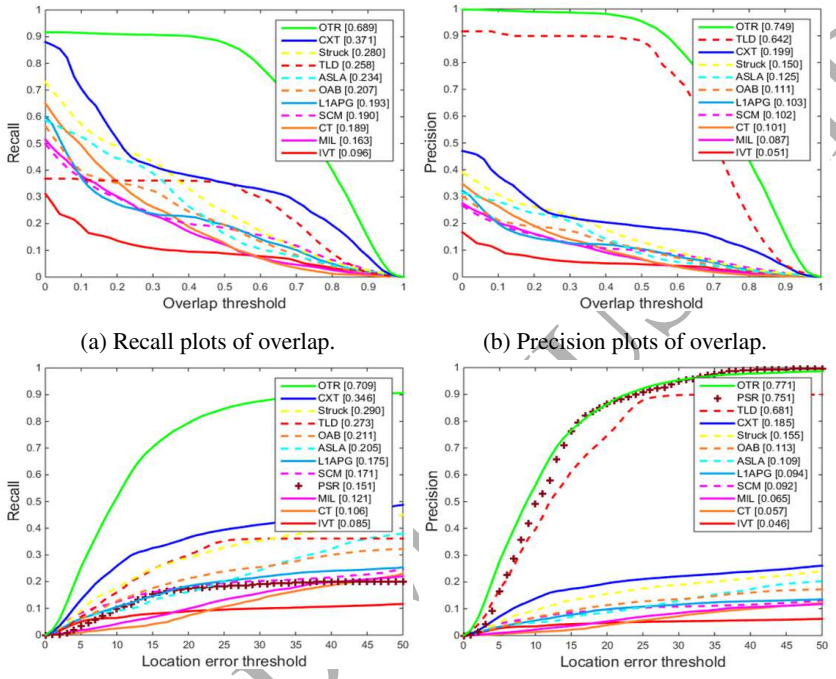
4.3. Comparison between online detectors

Before evaluating the retargeting results of the proposed framework, we compare our HRF with random ferns (Ozuysal et al., 2010), as well as our proposed detection cascade with the TLD detection cascade. Initially, we present the comparison of the proposed HRF and random ferns. It is well known that unlike offline training, online training is a challenging scenario because only a limited amount of samples are available for training. In this work, we have evaluated HRF and random ferns that are both trained in an online fashion, that is, training the classifiers using the samples generated in image t to detect the biopsy site in image $t + 1$. The aforementioned *in vivo* videos have been used in this study. The positive samples are retrieved by performing affine-warping on the ground truth at time t, and the negative samples are the high-variance image patches that do not overlap with the ground truth. We ensure that the samples used for training HRF and random ferns are the same, to enable a fair comparison.

For each image (after the first image where the optical biopsy site has been selected), multi-scale window scanning is used to generate around $300,000$ image patches, and the two classifiers are then used to rank these patches in a descending order according to their confidence scores. We define an image patch as a correct detection when its overlap ratio with the ground truth is larger than 0.5. Different parameter settings of M and Z have also been tested in our experiments. For HRF, M represents the number of rectangle sets and Z represents the number of rectangles in each set as mentioned. For random ferns, M represents the number of fern sets and Z represents the number of ferns in each set. These results have been provided in Fig. 5a for varying confidence thresholds θ, showing that the proposed HRF consistently outperforms random ferns.

Furthermore, experiments have been conducted to compare the proposed cascade with the TLD cascade. Precision-recall curves are generated for three different accuracy levels. Here, the accuracy is defined as the overlap ratio or the centre location distance between the results and the ground truth. The low, medium and high accuracy levels based on the overlap ratio correspond to the ratio thresholds equal to 0.3, 0.5 and 0.7, respectively. For location error, the low, medium and high accuracy levels correspond to the error thresholds equal to 40, 20, and 10 pixels, respectively.

For the precision and recall analysis, the detection results of the compared cascades have first been ranked according to the confidence of their outputs which in TLD cascade is the probability ratio, and in our cascade is the number of inliers from PROSAC. Then, the curves were generated by varying the thresholds of the probability ratio and the inlier number for TLD and OTR, respectively. The results are provided in Fig. 5b and Fig. 5c, and show that our detection cascade presents better results for all accuracy levels. For more detailed quantitative performance evaluation, we set the confidence threshold of TLD to 0.65 (the setting in the original paper), and set the threshold of inlier number of our method to 15. With these, the precision, recall and F-measure values based on the overlap ratio can be obtained and presented in Table 3, and indicate that our online detection cascade achieves better performance.



(a) Recall plots of overlap. (b) Precision plots of overlap.

(c) Recall plots of location error. (d) Precision plots of location error.

Figure 6: Plots of precision and recall values regarding varying overlap and centre location errors. For (a), (b), (d) and (e), the methods have been ranked and the performance score is defined as the area-under-curve (AUC) shown in the legend.

4.4. Retargeting results

Tissue tracking in *in vivo* GI videos is challenging, due to tissue deformation, fast camera motion, as well as specular highlights and artefacts caused by fluids. In addition, the operator often switches between imaging modes (NBI and normal white-light modes) during examinations. This can dramatically change the tissue appearance, resulting in frequent tracking failure. Furthermore, as the endoscopic camera has a small FOV, it is common that the selected biopsy site would disappear or re-enter the FOV after disappearance. In this section, we show the results of our proposed OTR (combined with the FB tracker) compared to the state-of-the-art online tracking methods.

4.4.1. Quantitative analysis

The overall results of the evaluated methods on the *in vivo* videos are presented in Figs. 6a, 6b, 6c and 6d, which verify that our approach (combined with the FB tracker) performs significantly better than the compared state-of-the-art trackers. This is mainly thanks to the online detetion cascade, which enables biopsy sites to be re-detected even after their disappearance. Most of the existing methods do not have an online detection component. It would be expected that TLD and CXT should have good performance as they incorporate online detectors. However, because their detectors adopt a nearest-neighbour classifier using normalised cross correlation for template matching, they are vulnerable when there are regions that have similar appearance to the tracked biopsy sites. In contrast, in our

Table 4: Quantitative results presented by $\frac{precision/recall}{F-measure}$ on 10 *in vivo* GI video sequences. **Bold** numbers represent the best performance, and *italic* numbers represent the second best performance. 'OTR–' denotes the OTR framework without performing location refinement.

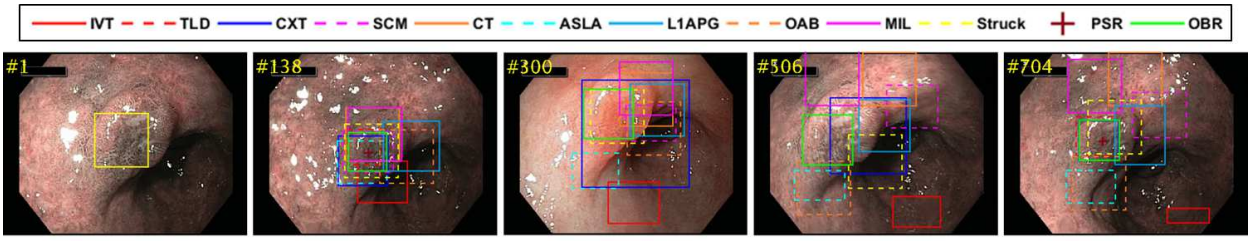
Seq.	IVT	TLD	CXT	SCM	CT	ASLA	L1APG	OAB	MIL	Struck	**OTR–**	**OTR**
#1	0.03/0.03 0.03	0.93/0.52 0.66	0.37/0.42 0.39	0.29/0.32 0.30	0.05/0.06 0.06	0.10/0.11 0.11	0.04/0.04 0.04	0.04/0.05 0.04	0.03/0.04 0.04	0.16/0.18 0.17	*0.91/0.89* *0.90*	**0.99/0.92** **0.96**
#2	0.08/0.16 0.11	0.88/0.31 0.46	0.20/0.37 0.26	0.09/0.16 0.11	0.04/0.08 0.06	0.09/0.16 0.11	0.13/0.24 0.17	0.08/0.16 0.11	0.04/0.07 0.05	0.09/0.16 0.11	*0.93/0.93* *0.93*	**0.96/0.96** **0.96**
#3	0.04/0.09 0.06	*0.99/0.73* *0.84*	0.28/0.44 0.26	0.11/0.27 0.16	0.05/0.13 0.08	0.06/0.14 0.08	0.04/0.09 0.06	0.06/0.14 0.08	0.03/0.08 0.05	0.06/0.15 0.09	0.74/0.72 0.73	**0.97/0.97** **0.97**
#4	0.13/0.23 0.17	1.00/0.59 0.74	0.44/0.75 0.55	0.13/0.23 0.17	0.07/0.12 0.09	0.13/0.22 0.16	0.09/0.16 0.12	0.22/0.37 0.27	0.05/0.09 0.06	0.13/0.22 0.17	*0.89/0.73* *0.80*	**1.00/0.87** **0.93**
#5	0.18/0.42 0.26	*0.89/0.88* *0.88*	0.38/0.86 0.52	0.18/0.41 0.25	0.10/0.23 0.14	0.19/0.43 0.26	0.19/0.43 0.26	0.18/0.40 0.25	0.18/0.41 0.25	0.18/0.42 0.26	0.80/0.80 0.80	**1.00/0.99** **0.99**
#6	0.02/0.03 0.02	1.00/0.00 0.01	0.02/0.03 0.02	0.03/0.07 0.04	0.24/0.47 0.32	0.02/0.04 0.02	0.05/0.09 0.06	0.21/0.40 0.27	0.01/0.02 0.02	0.29/0.57 0.38	*0.60/0.72* *0.65*	**1.00/0.64** **0.78**
#7	0.00/0.00 0.00	1.00/0.00 0.00	0.05/0.06 0.05	0.06/0.07 0.06	0.06/0.07 0.06	0.14/0.16 0.15	0.25/0.29 0.27	0.16/0.18 0.17	0.14/0.16 0.15	0.27/0.31 0.29	*0.53/0.40* *0.46*	**0.89/0.87** **0.88**
#8	0.02/0.04 0.03	0.31/0.04 0.08	0.03/0.07 0.04	0.02/0.04 0.03	0.08/0.17 0.10	0.07/0.16 0.10	0.02/0.05 0.03	0.08/0.18 0.11	0.01/0.02 0.01	0.04/0.09 0.05	*0.58/0.26* *0.36*	**0.99/0.62** **0.76**
#9	0.05/0.11 0.07	1.00/0.20 0.34	0.44/0.95 0.61	0.09/0.20 0.13	0.18/0.40 0.25	0.05/0.11 0.07	0.29/0.61 0.39	0.28/0.60 0.38	0.21/0.45 0.29	0.29/0.61 0.39	*0.96/0.73* *0.83*	**1.00/0.91** **0.95**
#10	0.03/0.14 0.05	*0.58/0.66* *0.62*	0.16/0.82 0.26	0.03/0.17 0.05	0.03/0.18 0.06	0.03/0.17 0.05	0.03/0.18 0.06	0.04/0.22 0.07	0.03/0.18 0.06	0.04/0.18 0.06	0.92/0.39 0.55	**0.99/0.60** **0.75**
Overall	0.05/0.09 0.06	0.88/0.35 0.51	0.18/0.35 0.24	0.10/0.18 0.13	0.07/0.13 0.09	0.08/0.16 0.11	0.10/0.20 0.13	0.10/0.20 0.13	0.06/0.12 0.08	0.13/0.24 0.17	*0.75/0.64* *0.69*	**0.95/0.88** **0.91**

Table 5: Quantitative results presented by average centre location errors (in pixels) on 10 *in vivo* GI video sequences. **Bold** numbers represent the best performance, and *italic* numbers represent the second best performance. Failures are denoted as '-'. 'OTR–' denotes the OTR framework without performing location refinement.

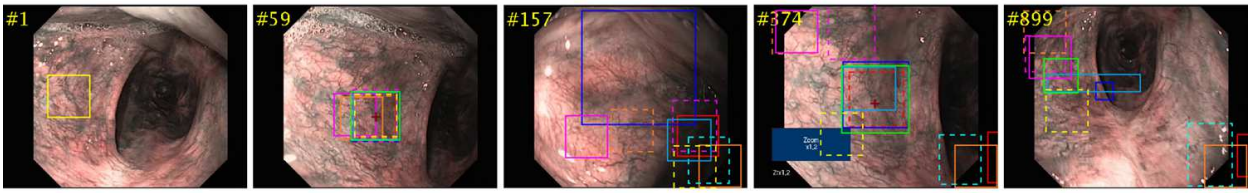
Seq.	Images	IVT	TLD	CXT	SCM	CT	ASLA	L1APG	OAB	MIL	Struck	**PSR**	**OTR–**	**OTR**
#1	705	210	20	46	122	132	116	137	127	126	98	33	*14*	**8**
#2	1003	354	**4**	81	107	308	249	74	204	112	95	*7*	10	8
#3	1700	77	*12*	150	141	55	122	150	128	128	41	**10**	18	**10**
#4	336	319	*12*	34	151	159	232	265	74	173	126	15	14	**8**
#5	493	66	*13*	38	70	137	137	176	105	112	80	16	17	**11**
#6	325	126	-	145	91	80	108	201	110	192	107	-	*9*	**7**
#7	1349	287	-	81	249	108	52	126	246	215	70	-	*34*	**19**
#8	578	194	69	102	159	199	123	334	57	156	192	21	*15*	**10**
#9	266	94	**7**	11	85	52	87	94	19	49	79	*8*	12	**7**
#10	1139	451	14	28	270	131	159	254	174	265	279	*10*	19	**9**
Overall	7894	230	*14*	82	165	136	124	155	160	160	96	*14*	19	**12**

proposed detector, PROSAC has been included to verify the local shape context of the candidate regions, which provides more robust detection against false positives. Compared to our previous PSR method, OTR has achieved competitive performance in precision values, and significantly better results in recall values in terms of location error, which corresponds to the issue of planar assumption of PSR.

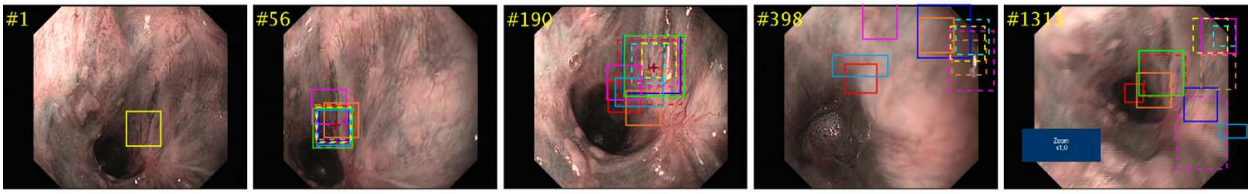
To further evaluate the performance of the proposed algorithm, we set the threshold of the overlap ratio to 0.5 and present the precision, recall and F-measure values in Table 4. All the methods can be ranked using the F-measures, and the best and second best approaches for each sequence have been highlighted. To show the effect of location refinement, we have also presented the results of 'OTR–' which is the OTR framework without performing location refinement. Overall, it can been seen that our approach OTR provides the highest F-measure 0.91, while TLD is ranked second with F-measure equal to 0.51. We can also observe that most approaches (IVT, OAB, MIL, Struck, etc.) presented very low F-measures for all of the GI sequences. This is because the biopsy sites are not always in the FOV of the camera, which is against their assumption of the consistent appearance of an object. Moreover, the detectors of TLD and PSR are sensitive to the initial biopsy site locations, and this has resulted in the failures in Seq. 6 and Seq. 7 where the biopsy sites were located on regions that have low intensity variances.



(a) Seq. 1 with the imaging mode switched between NBI and white-light.



(b) Seq. 2 with the biopsy site out of FOV in image #157.



(c) Seq. 3 with the biopsy site out of FOV in image #398.

Figure 7: Snapshots of video results of Seqs. 1-3. The first images of each row display the selected optical biopsy site location. Best viewed on colour display.

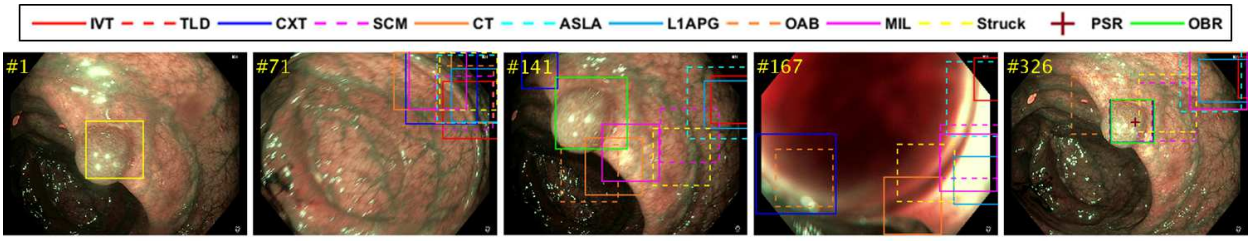
We have also presented the evaluation results using average centre location errors for each sequence. As the size of a biopsy site would change in the camera view due to endoscopic movement, all the location errors have been re-scaled to the first image where the site is selected. These errors are provided in Table 5. It has been observed that TLD and PSR failed in Seq. 6 and Seq. 7, where it could not present tracking output more than 10 images after the biopsy sites were selected. The overall location errors have demonstrated that the proposed OTR framework presented the best accuracy among all the methods, with an overall error of 12 pixels. The main reason for this good performance is that biopsy site locations are refined at the final stage of our framework using the inliers of keypoint correspondences identified by shape context.

4.4.2. Qualitative analysis

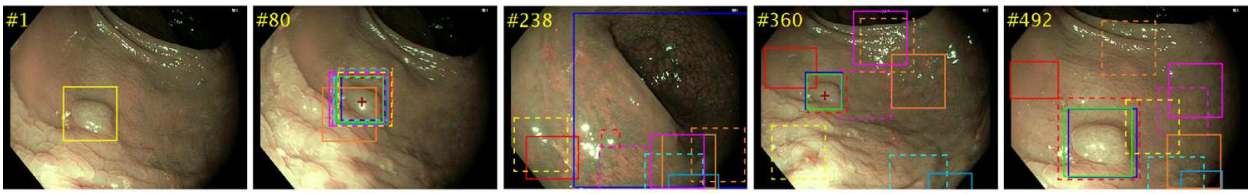
Robustness to small FOV. As mentioned above, the optical biopsy sites often move in and out of the FOV during the examination. Retargeting is required when the camera moves back to the same biopsy sites. As shown in Fig. 7 and Fig. 8, many of the compared trackers do not have a verification component to identify the tracking failures, which has led to tracking wrong regions when the biopsy site is out of FOV, for instance, image #157 in Seq. 2, image #398 in Seq. 3, and images #71 and #167 in Seq. 4. The trackers that failed include IVT, CXT, SCM, CT, ASLA, L1APG, OAB, MIL and Struck.

Changing of imaging modes. During an endoscopic procedure, the operator can switch between different imaging modes for scrutinizing suspicious anatomical areas of interest. After switching the mode, the appearance of the tissue can change dramatically. As shown in image #300 in Fig. 7a and image #1 in Fig. 9d, the white-light images have less vascular features than the NBI images. Nevertheless, due to the structured SVM applied in our framework, the appearance of the keypoints can be updated online. This enables our method to re-detect the biopsy site even when the imaging mode has been changed (Fig. 7a and Fig. 9d).

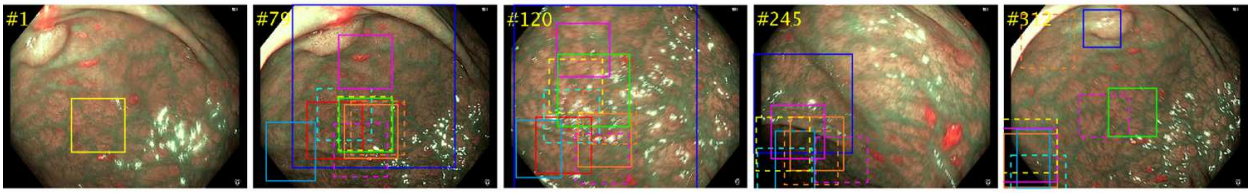
Tissue deformation. Tissue deformation is a challenging issue in medical image analysis. During GI examina-



(a) Seq. 4 with the biopsy site out of FOV in image #71 and #167.



(b) Seq. 5 with the biopsy site out of FOV in image #238.



(c) Seq. 6 with the biopsy site out of FOV in image #245.

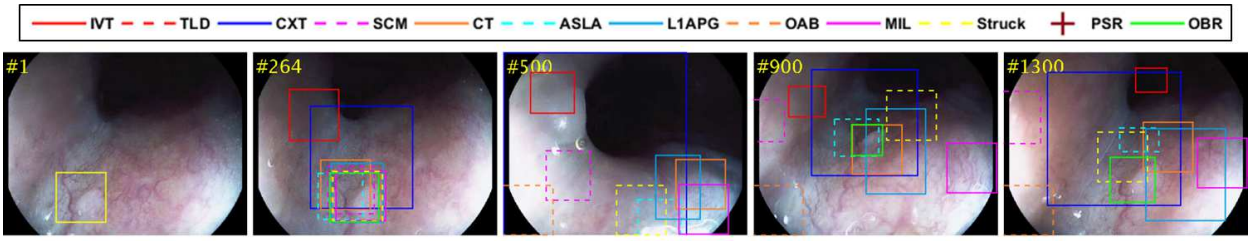
Figure 8: Snapshots of video results of Seqs. 4-6. The first images of each row display the selected optical biopsy site location. Best viewed on colour display.

tions, global deformation occurs caused by patient movement, and peristalsis or respiratory induced motion. During optical biopsy procedures, the sites of interest are usually the local regions, thus they would not be affected significantly by global tissue deformation. The robustness of our approach to global deformation has been shown on the *in vivo* experiments, as can be seen in Figs. 7, 8 and 9 and supplementary videos.

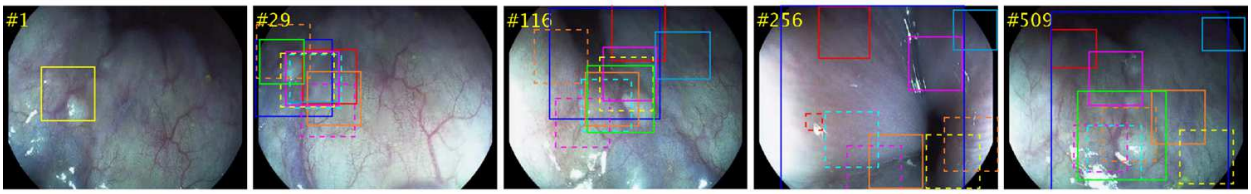
Specular highlights. As the endoscopic camera is close to the tissue surface during navigation, the presence of specular highlights on the images cannot be ignored. It can be seen, for example, in Figs. 7a, 8a and 8b that these specularities can cause occlusions at the biopsy sites. In our framework, the HRF classifier and shape context consider partial information (local region comparison and keypoints) of the biopsy site, which has shown good performance in Seq. 4 and Seq. 5 where specular occlusions exist.

Robustness to false positives. One common issue of the online detectors is the limited number of samples available for training the classifiers, leading to under-fitting. This is typically manifested as the poor performance of the classifier in distinguishing true positives from false negatives. As presented in images #238 (Fig. 8b), #256 (Fig. 9b), #80 and #919 (Fig. 9d), the detectors of CXT and TLD have generated false positives that have similar appearance to the true biopsy sites. In contrast, the proposed OTR is robust to these errors, thanks to the use of shape context for verification.

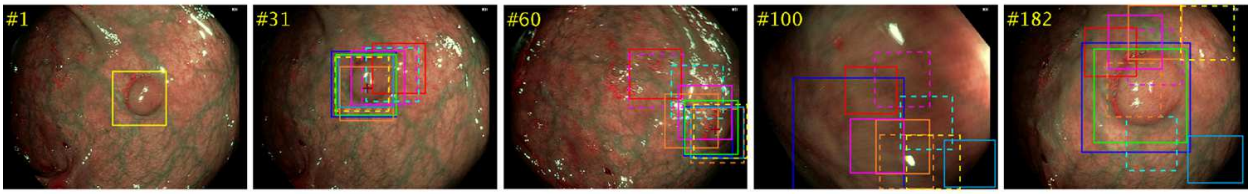
Rotation and scale changes. As we have generated affine-warped positive samples for updating the HRF, and performed the image scanning with varying scales, our framework is robust to in-plane rotation and scale variations of the biopsy sites. The robustness of the proposed framework to these can be observed from the qualitative results (Figs. 7-9) and the supplementary videos. The *in vivo* GI videos used for validation were collected during standard endoscopic procedures with varying scale and rotation conditions.



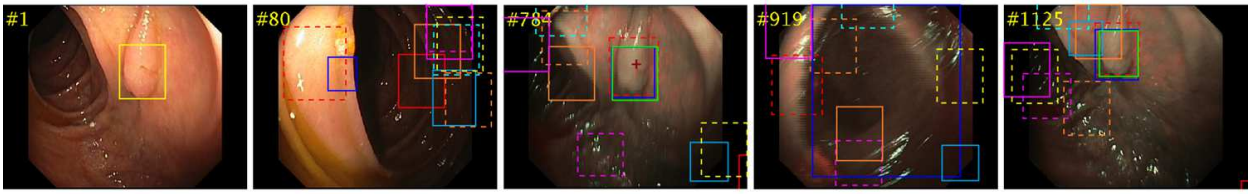
(a) Seq. 7 with the biopsy site out of FOV in image #500.



(b) Seq. 8 with the biopsy site out of FOV in image #256.



(c) Seq. 9 with the biopsy site out of FOV in image #100.



(d) Seq. 10 with the biopsy site out of FOV in image #80 and #919 and imaging mode switched between NBI and white-light.

Figure 9: Snapshots of video results of Seqs. 7-10. The first images of each row display the selected optical biopsy site location. Best viewed on colour display.

5. Conclusions and future works

In this paper, we have addressed optical biopsy retargeting as a tracking-by-detection task. An online traking and retargeting framework termed OTR has been proposed to re-detect optical biopsy sites. A random binary descriptor based on Haar-like features has been introduced based on local region comparisons, which can be formulated as a simple random forest classifier. To enable robust retargeting, a novel cascaded detector has been introduced which incorporates an online random forest classifier, a structured SVM and a verification component. The results derived have shown that, the proposed detection verification approach, based on a novel combination of shape context and PROSAC, significantly improves the robustness of the detection to false positives. In addition, the proposed cascade can be easily combined with other temporal trackers, and thus is of generic value for other tracking applications. The online framework runs in real-time, enabling its practical use in a clinical set-up. Quantitative and qualitative performance evaluation has been conducted on a large dataset of challenging *in vivo* GI video sequences, with results demonstrating that our proposed framework outperforms the current state-of-the-art in terms of both accuracy and robustness. To facilitate benchmarking of tissue tracking techniques, the *in vivo* GI dataset used, along with the ground truth, are made available online.

It should be noted that in this paper, our implementation only considers retargeting of one optical biopsy site.

For simultaneous retargeting of multiple sites, implementation optimisation on existing multi-core hardware can be performed. For further performance enhancement in such cases, joint training of random Haars with structured SVMs can be explored to share common computational workloads. Clinically, one important extension of the proposed framework is to use it for serial examinations of patient. In these cases, large appearance differences would be encountered especially for patients undergoing chemo- or radio-therapy. How to effectively link the trained classifiers for inter-procedure retargeting with due consideration of global scene context (Ye et al., 2014) would be an important future research topic.

Appendix A.

Table A.1: Notations of main mathematical terms.

Notation	Definition	Notation	Definition
M	Number of rectangle sets	$\mathbf{w}_t$	The weight vector at time t
Z	Number of rectangles in each set	$\xi_{t,i}$	The slack variable at t for image patch i
$\mathbf{d}$	Binary code sets and a feature descriptor	K	Number of top candidates stored in $\hat{\mathcal{Y}}$
d_m	m-th binary code in $\mathbf{d}$	$\hat{\mathbf{y}}_{t,k}$	k-th candidate in estimation results $\hat{\mathcal{Y}}_t$
$\mathcal{X}$	A sequence of images	$\mathbf{v}$	The weight vector in structured SVM
$\mathbf{x}$	An image in $\mathcal{X}$	$\mathbf{v}_i$	i-th segment in $\mathbf{v}$
$\mathcal{Y}$	A set of image patches	$\mathbf{v}_t$	The weight vector $\mathbf{v}$ at time t
$\mathbf{y}$	An image patch in $\mathcal{Y}$	$\mathbf{v}_{t,i}$	i-th segment in $\mathbf{v}$ at time t
p	Frequency of an integer number in the positive histogram	$\mathbf{h}$	A transform
n	Frequency of an integer number in the negative histogram	$\mathbf{h}_t^\star$	The best transform at time t
θ	The threshold to posterior probability	$\mathbf{h}_t'$	The second best transform at time t
$\hat{\mathbf{y}}$	An estimation result	$\mathbf{c}_t$	A correspondence set of keypoints at time t
$\hat{\mathcal{Y}}$	A set of estimation results	$\mathbf{c}_t^\star$	The best correspondence set at time t
$\hat{\mathcal{Y}}_t$	A set of estimation results at time t	$\mathbf{c}_{t,k}$	k-th correspondence set at time t
$\mathbf{w}$	The weight vector in structured SVM	I	Number of model keypoints
λ	A constant scaling parameter in structured SVM	u_i	A model keypoint with index i
$\mathbf{x}_t$	Image at time t	v_j	A keypoint with index j corresponding to u_i
$\mathbf{y}_{t,i}$	Image patch i at time t	$\mathbf{d}_j^\star$	The descriptor of the best match to u_i
$\mathbf{y}_t^\star$	The best estimation result at time t	$\mathbf{d}_j'$	The descriptor of the second best match to u_i
r_t^x	The x coordinate of $\mathbf{y}_t^\star$	T_c	The threshold to the number of inliers in PROSAC
r_t^y	The y coordinate of $\mathbf{y}_t^\star$	$\mathcal{S}$	A set of scale factors
r_t^w	The width of $\mathbf{y}_t^\star$	s	A scale factor
r_t^h	The height of $\mathbf{y}_t^\star$	$s^\star$	The best scale factor in $\mathcal{S}$
η_t	The step size at time t in structure SVM		

References

Allain, B., Hu, M., Lovat, L.B., Cook, R.J., Vercauteren, T., Ourselin, S., Hawkes, D.J., 2012. Re-localisation of a biopsy site in endoscopic images and characterisation of its uncertainty. Medical Image Analysis 16, 482 – 496.

Atasoy, S., Glocker, B., Giannarou, S., Mateus, D., Meining, A., Yang, G.Z., Navab, N., 2009. Probabilistic Region Matching in Narrow-Band Endoscopy for Targeted Optical Biopsy, in: Yang, G.Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2009. Springer Berlin Heidelberg. volume 5761 of *Lecture Notes in Computer Science*, pp. 499–506.

Atasoy, S., Mateus, D., Meining, A., Yang, G.Z., Navab, N., 2012. Endoscopic Video Manifolds for Targeted Optical Biopsy. Medical Imaging, IEEE Transactions on 31, 637–653.

Babenko, B., Yang, M.H., Belongie, S., 2011. Robust Object Tracking with Online Multiple Instance Learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on 33, 1619–1632.

Bao, C., Wu, Y., Ling, H., Ji, H., 2012. Real time robust l1 tracker using accelerated proximal gradient approach, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1830–1837.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. SURF: Speeded Up Robust Features, in: Leonardis, A., Bischof, H., Pinz, A. (Eds.), Computer Vision ECCV 2006. Springer Berlin Heidelberg. volume 3951 of *Lecture Notes in Computer Science*, pp. 404–417.

Belongie, S., Malik, J., Puzicha, J., 2002. Shape matching and object recognition using shape contexts. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 509–522.

Bertelli, L., Yu, T., Vu, D., Gokturk, B., 2011. Kernelized structural SVM learning for supervised object segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 2153–2160.

Blaschko, M., Lampert, C., 2008. Learning to Localize Objects with Structured Output Regression, in: Forsyth, D., Torr, P., Zisserman, A. (Eds.), Computer Vision ECCV 2008. Springer Berlin Heidelberg. volume 5302 of *Lecture Notes in Computer Science*, pp. 2–15.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P., 2012. BRIEF: Computing a Local Binary Descriptor Very Fast. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, 1281–1298.

Chum, O., Matas, J., 2005. Matching with PROSAC - progressive sample consensus, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, pp. 220–226 vol. 1.

Dinh, T.B., Vo, N., Medioni, G., 2011. Context tracker: Exploring supporters and distracters in unconstrained environments, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 1177–1184.

Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A., 2010. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision 88, 303–338.

Giannarou, S., Visentini-Scarzanella, M., Yang, G.Z., 2013. Probabilistic Tracking of Affine-Invariant Anisotropic Regions. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35, 130–143.

Grabner, H., Grabner, M., Bischof, H., 2006. Real-Time Tracking via On-line Boosting, in: Proceedings of the British Machine Vision Conference, BMVA Press. pp. 6.1–6.10.

Hare, S., Saffari, A., Torr, P., 2011. Struck: Structured output tracking with kernels, in: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 263–270.

Hare, S., Saffari, A., Torr, P., 2012. Efficient online structured output learning for keypoint-based object tracking, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1894–1901.

Hughes, M., Yang, G.Z., 2012. Robotics and smart instruments for translating endomicroscopy to in situ, in vivo applications. Computerized Medical Imaging and Graphics 36, 589–590.

Jia, X., Lu, H., Yang, M.H., 2012. Visual tracking via adaptive structural local sparse appearance model, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1822–1829.

Kalal, Z., Mikolajczyk, K., Matas, J., 2010. Forward-Backward Error: Automatic Detection of Tracking Failures, in: Pattern Recognition (ICPR), 2010 20th International Conference on, pp. 2756–2759.

Kalal, Z., Mikolajczyk, K., Matas, J., 2012. Tracking-Learning-Detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, 1409–1422.

Leutenegger, S., Chli, M., Siegwart, R., 2011. BRISK: Binary Robust Invariant Scalable Keypoints, in: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2548–2555.

Mountney, P., Giannarou, S., Elson, D., Yang, G.Z., 2009. Optical Biopsy Mapping for Minimally Invasive Cancer Screening, in: Yang, G.Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2009. Springer, Heidelberg. volume 5761 of *Part I. LNCS*, pp. 483–490.

Mountney, P., Lo, B., Thiemjarus, S., Stoyanov, D., Zhong-Yang, G., 2007. A Probabilistic Framework for Tracking Deformable Soft Tissue in Minimally Invasive Surgery, in: Ayache, N., Ourselin, S., Maeder, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2007. Springer Berlin Heidelberg. volume 4792 of *Lecture Notes in Computer Science*, pp. 34–41.

Mountney, P., Stoyanov, D., Yang, G.Z., 2010. Three-Dimensional Tissue Deformation Recovery and Tracking. Signal Processing Magazine, IEEE 27, 14–24.

Mountney, P., Yang, G.Z., 2008. Soft Tissue Tracking for Minimally Invasive Surgery: Learning Local Deformation Online, in: Metaxas, D., Axel, L., Fichtinger, G., Szkely, G. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2008. Springer Berlin Heidelberg. volume 5242 of *Lecture Notes in Computer Science*, pp. 364–372.

Mountney, P., Yang, G.Z., 2010. Motion Compensated SLAM for Image Guided Surgery, in: Jiang, T., Navab, N., Pluim, J., Viergever, M. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2010. Springer Berlin Heidelberg. volume 6362 of *Lecture Notes in Computer Science*, pp. 496–504.

Nebehay, G., Pflugfelder, R., 2014. Consensus-based matching and tracking of keypoints for object tracking, in: Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, pp. 862–869.

Newton, R., Kemp, S., Shah, P., Elson, D., Darzi, A., Shibuya, K., Mulgrew, S., Yang, G.Z., 2011. Progress Toward Optical Biopsy: Bringing the Microscope to the Patient. Lung 189, 111–119.

Newton, R., Noonan, D., Vitiello, V., Clark, J., Payne, C., Shang, J., Sodergren, M., Darzi, A., Yang, G.Z., 2012. Robot-assisted transvaginal peritoneoscopy using confocal endomicroscopy: a feasibility study in a porcine model. Surgical Endoscopy 26, 2532–2540.

Oron, S., Bar-Hillel, A., Levi, D., Avidan, S., 2012. Locally Orderless Tracking, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1940–1947.

Ozuysal, M., Calonder, M., Lepetit, V., Fua, P., 2010. Fast Keypoint Recognition Using Random Ferns. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, 448–461.

Reiter, A., Allen, P.K., Zhao, T., 2014. Appearance learning for 3d tracking of robotic surgical tools. The International Journal of Robotics Research 33, 342–356.

Ren, X., Malik, J., 2003. Learning a classification model for segmentation, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pp. 10–17 vol.1.

Richa, R., B, A.P., Poignet, P., 2011. Towards robust 3d visual tracking for motion compensation in beating heart surgery. Medical Image Analysis 15, 302 – 315.

Richa, R., Vgvlgyi, B., Balicki, M., Hager, G., Taylor, R., 2012. Hybrid Tracking and Mosaicking for Information Augmentation in Retinal Surgery, in: Ayache, N., Delingette, H., Golland, P., Mori, K. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI

2012. Springer Berlin Heidelberg. volume 7510 of *Lecture Notes in Computer Science*, pp. 397–404.

Ross, D., Lim, J., Lin, R.S., Yang, M.H., 2008. Incremental Learning for Robust Visual Tracking. International Journal of Computer Vision 77, 125–141.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF, in: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2564–2571.

Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A., 2011. Pegasos: primal estimated sub-gradient solver for SVM. Mathematical Programming 127, 3–30.

Sznitman, R., Ali, K., Richa, R., Taylor, R., Hager, G., Fua, P., 2012. Data-Driven Visual Tracking in Retinal Microsurgery, in: Ayache, N., Delingette, H., Golland, P., Mori, K. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2012. Springer Berlin Heidelberg. volume 7511 of *Lecture Notes in Computer Science*, pp. 568–575.

Sznitman, R., Becker, C., Fua, P., 2014. Fast Part-Based Classification for Instrument Detection in Minimally Invasive Surgery, in: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2014. Springer International Publishing. volume 8674 of *Lecture Notes in Computer Science*, pp. 692–699.

Sznitman, R., Richa, R., Taylor, R., Jedynak, B., Hager, G., 2013. Unified Detection and Tracking of Instruments during Retinal microsurgery. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35, 1263–1273.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features, in: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, pp. I–511–I–518 vol.1.

Wu, Y., Lim, J., Yang, M.H., 2013. Online Object Tracking: A Benchmark, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 2411–2418.

Yao, R., Shi, Q., Shen, C., Zhang, Y., van den Hengel, A., 2012. Robust Tracking with Weighted Online Structured Learning, in: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.), Computer Vision ECCV 2012. Springer Berlin Heidelberg. volume 7574 of *Lecture Notes in Computer Science*, pp. 158–172.

Ye, M., Giannarou, S., Patel, N., Teare, J., Yang, G.Z., 2013. Pathological Site Retargeting under Tissue Deformation Using Geometrical Association and Tracking, in: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2013. Springer Berlin Heidelberg. volume 8150, pp. 67–74.

Ye, M., Johns, E., Giannarou, S., Yang, G.Z., 2014. Online Scene Association for Endoscopic Navigation, in: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2014. Springer International Publishing. volume 8674 of *Lecture Notes in Computer Science*, pp. 316–323.

Zhang, K., Zhang, L., Yang, M.H., 2012. Real-Time Compressive Tracking, in: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.), Computer Vision ECCV 2012. Springer Berlin Heidelberg. volume 7574 of *Lecture Notes in Computer Science*, pp. 864–877.

Zhong, W., Lu, H., Yang, M.H., 2012. Robust object tracking via sparsity-based collaborative model, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1838–1845.