Online Scene Association for Endoscopic Navigation

Menglong Ye[1], Edward Johns[2], Stamatia Giannarou[1] and Guang-Zhong Yang[1]

[1]The Hamlyn Centre for Robotic Surgery, Imperial College London, UK
[2]Department of Computer Science, University College London, UK
menglong.ye11@imperial.ac.uk

Abstract. Endoscopic surveillance is a widely used method for monitoring abnormal changes in the gastrointestinal tract such as Barrett's esophagus. Direct visual assessment, however, is both time consuming and error prone, as it involves manual labelling of abnormalities on a large set of images. To assist surveillance, this paper proposes an online scene association scheme to summarise an endoscopic video into scenes, on-the-fly. This provides scene clustering based on visual contents, and also facilitates topological localisation during navigation. The proposed method is based on tracking and detection of visual landmarks on the tissue surface. A generative model is proposed for online learning of pairwise geometrical relationships between landmarks. This enables robust detection of landmarks and scene association under tissue deformation. Detailed experimental comparison and validation have been conducted on *in vivo* endoscopic videos to demonstrate the practical value of our approach.

1 Introduction

Gastrointestinal (GI) endoscopy is widely used for screening abnormal changes in the digestive tract. One of the major diseases in the digestive tract is Barrett's Esophagus (BE), which refers to metaplasia on the esophageal mucosa resulting from chronic gastroesophagel reflux, which has been regarded as a strong factor of esophageal adenocarcinoma. To monitor abnormal changes in BE, surveillance endoscopy has been a popular method. This involves post-processing of the endoscopic videos by expert pathologists, followed by serial examinations of the same patient. However, surveillance endoscopy can be complicated, as the post-processing stage often involves manual assessment and labelling of abnormalities on a large number of images, which is both time consuming and error prone. In cases when online retargeting is required for optical biopsy, the procedure becomes technically even more challenging [1].

In this work, we address the above issues by using endoscopic scene association, which refers to associating an endoscopic image to previously viewed scenes. In computer vision, scene association is the task of learning the visual information of a scene from a set of training data, followed by recognising a query image from candidate scenes. In endoscopic procedures, such as GI endoscopy,

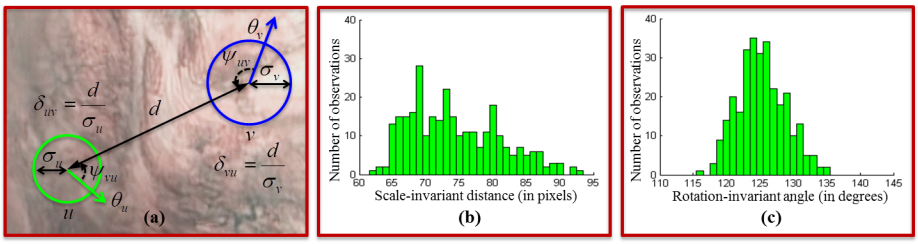

Fig. 1. (a) Pairwise relationships between two feature observations; Example distributions of (b) Scale-Invariant Distance (SID) and (c) Rotation-Invariant Angle (RIA) between two landmarks.

scene association can enable users to topologically localise the endoscopic camera position. This is particularly useful, for example, during probe-based confocal laser endomicroscopic (pCLE) procedures to facilitate retargeting of pathological sites. In addition, the derived scene clusters can be used to assist surveillance in follow-up examinations.

To achieve endoscopic scene association, dimensionality-reduction based on manifold learning [2] and semantic encoding [3] have been proposed. Feature-based methods such as Bag-of-Words (BoW) have been used to provide an atlas for confocal image retrieval [4]. These methods can learn the visual properties of particular scenes, but mostly are based on offline processing.

In this paper, we propose an online approach for endoscopic scene association based on robust visual tracking and detection. The goal of our approach is to summarise an endoscopic video into scenes, on-the-fly. Inspired by landmark-based recognition of [5,6], our approach samples visual landmarks on the tissue surface by analysing the stability of local features. To associate the current image to previously viewed scenes, an appearance-based cascaded classification scheme is adopted, together with matching to generative models of pairwise landmark geometries. These local pairwise relationships enable robustness to tissue deformation, which is not available in standard image-to-image matching. Our approach has been compared with commonly used image matching methods on *in vivo* GI studies, and results demonstrate the clinical value of our method.

2 Methods

2.1 Feature Tracking for Landmark Sampling

The proposed framework for scene association is based on the tracking of SIFT features [7] to sample landmarks on the tissue surface. We define landmarks as physical locations on the tissue surface, and features as the observations of landmarks. A landmark is represented by its set of features detected over an image sequence. Landmark geometry is then described as a distribution learned from the individual feature geometries.

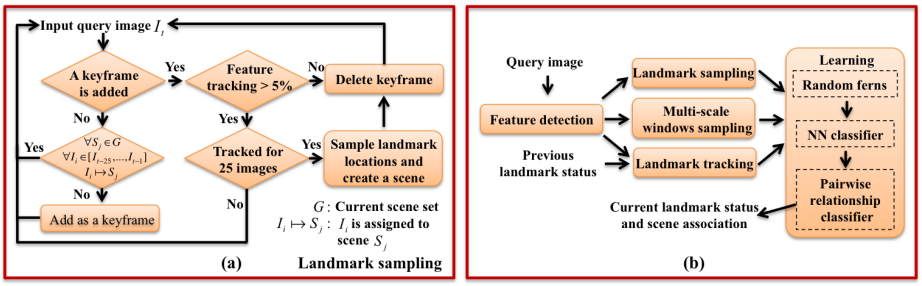

Fig. 2. (a) The flow chart of landmark sampling; (b) The overall framework diagram of the proposed scene association approach.

In practice, long-term tissue tracking is challenging in endoscopic navigation, due to tissue deformation, fast camera motion, and lighting variations. We propose a keyframe-based approach for feature tracking using optical flow enhanced with Forward-Backward (FB) error detection [8]. When a new keyframe is added (see Section 2.4), features are then tracked over an episode of one second (typically 25 frames), rejecting unstable features with low local intensity contrast. At the end of the episode, a new scene is created, with each stable feature track forming one landmark in that scene. In order to achieve long-term landmark recognition, a cascaded appearance classification scheme [9] has been adopted, which includes Random Ferns (RF) [10] and a nearest neighbour (NN) classifier. To detect a landmark in a query image, multi-scale windows are sampled at feature locations, which are then classified using RF. Candidates passing this stage are then classified by calculating Normalised Cross Correlation (NCC) and finding the NN to all previous features for this landmark. The initialisation and online updating process of RF and the NN classifier is similar to [1, 9].

With this online learning scheme, the method is able to re-identify landmarks when they re-enter the field-of-view (FOV) even if optical flow fails. However, the main limitation of these classifiers is that they fail to distinguish landmarks with similar appearance which can lead to tracking false positives, and updating the classifiers with incorrect samples.

2.2 Learning Pairwise Relationships

To deal with the limitation of appearance-only learners, we include a third classification component that incorporates scene geometry. Due to the deformable and non-affine nature of endoscopic environments, standard 3D image matching techniques such as estimating the fundamental matrix or homography [11] are not suitable. Rather than calculating a global scene alignment, we therefore focus on modelling pairwise relationships between landmarks, allowing for deformations and non-affine properties to be learned independently and locally. These relationships are defined as the Scale-Invariant Distance (SID) and Rotation-Invariant Angle (RIA). Each landmark pair has a distribution over these two

geometries, learned from the tracked features for those landmarks.

As shown in Fig. 1 (a), features u and v have orientations θ_u and θ_v, and sizes σ_u and σ_v. The distance in pixels between u and v in the image is denoted d. We then define SID from u to v as $\delta_{uv} = \frac{d}{\sigma_u}$, and define RIA ψ_{uv} from u to v as the angle from the feature orientation axis of v to the line connecting the feature locations. Similarly, the SID and RIA from v to u are $\delta_{vu} = \frac{d}{\sigma_v}$ and ψ_{vu}. These two measures have been used effectively in [6] and learned in an offline framework, however, we formulate this in an online model. With a set of images that contain both landmarks x and y, we can now obtain distributions of SID and RIA for the pair which we model as histograms in discrete space. Example distributions are shown in Fig. 1 (b) and (c).

2.3 Landmark Recognition

During recognition, given a feature u which has been assigned to landmark x from the appearance classifier, let us consider the likelihood that this is a true observation of x, by incorporating geometry. This likelihood is evaluated as a score $A(x)$, by taking all neighbouring landmarks $y \in \mathcal{Y}$ in the same scene as x, and averaging over the likelihoods from each of the pairs (x, y). If a_u denotes the appearance of u, and g_{uv} denotes the pairwise geometry of u and v, where v has been assigned to y from the appearance classifier, we define this score as:

$$A\left(x\right) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} p\left(x|a_u, g_{uv}\right), \tag{1}$$

where the likelihood from each pair is evaluated using Bayes' rule:

$$p\left(x|a_u, g_{uv}\right) = \frac{p(g_{uv}|x)p(x|a_u)}{p(g_{uv}|x)p(x|a_u) + p(g_{uv}|\bar{x})p(\bar{x}|a_u)}. \tag{2}$$

Here, $p\left(x|a_u\right)$ is the confidence score from the appearance classifier, and $p\left(\bar{x}|a_u\right)$ is the score of a false positive appearance classification, set to $1 - p\left(x|a_u\right)$.

Let us now denote δ_{ab} and ψ_{ab} as the SID and RIA respectively from a to b. In the case where a and b are features, δ_{ab} and ψ_{ab} are the observed geometries in the current image, and in the case of landmarks, they are the learned distributions over geometries from training. The observation likelihood of the pairwise geometry of u and v is then defined as:

$$p(g_{uv}|x) = p\left(\delta_{uv}|\delta_{xy}\right) p\left(\psi_{uv}|\psi_{xy}\right) p\left(y|x\right), \tag{3}$$

where $p\left(y|x\right)$ is the observation rate of y in images containing x, and $p\left(\delta_{uv}|\delta_{xy}\right)$ and $p\left(\psi_{uv}|\psi_{xy}\right)$ are taken from the associated SID and RIA distributions for the landmark pair, respectively.

Returning to Eq. 2, the likelihood of observing the pairwise geometry given a false positive appearance classification of x, is defined as:

$$p\left(g_{uv}|\bar{x}\right) = p\left(\delta_{uv}\right) p\left(\psi_{uv}\right), \tag{4}$$

where $p\left(\delta_{uv}\right)$ and $p\left(\psi_{uv}\right)$ are the priors of randomly observing these geometries taken over the full distribution across all landmark pairs in the sequence.

2.4 Online Learning and Scene Association

Each scene is initialised using an episode of images to track robust features, build the landmarks, and learn the initial pairwise landmark geometry distributions. If a landmark detected in a query image satisfies $A(x) > T_l$, it is then tracked over subsequent frames using FB tracking on square regions centred at the landmark (similar to the feature tracking in Section 2.1). These tracked features update the landmark appearance, $p(y|x)$, and the SID and RIA distributions.

We define a score for assigning the query image to existing scene S, with its landmark set $\mathcal{X}$, as

$$B\left(S\right) = \frac{\sum_{x\in\mathcal{X}} A\left(x\right)}{\sum_{x\in\mathcal{X}} p\left(x\right)}, \tag{5}$$

where $p\left(x\right)$ is the observation rate of x in the scene and normalises for scenes with varying numbers of landmarks. The scene label S^* of a query image is the scene with the greatest score, and we accept the assignment when $B(S^*) > T_s$.

For online learning, we introduce a new scene when the endoscope has moved sufficiently, based on the following criteria (see Fig. 2 (a)): given image I_t, the images over the last second are checked $[I_{t-25}...I_{t-1}]$, and I_t is added as a keyframe only when none of these images was assigned to a previously viewed scene; after the keyframe is added, landmark sampling is performed from I_t to I_{t+25}, however, it is terminated whenever the successfully tracked features are below 5% of the total number of features detected in the keyframe. The advantages of these criteria are twofold. Firstly, the overlap area between two scenes is minimised; secondly, the uninformative images caused by motion blur or other artefacts are filtered out. Fig. 2 (b) shows the overall framework of the proposed online scene association approach.

Parameter Smoothing. If we record only the observed, explicit distributions of SIDs and RIAs, overfitting can cause false negative landmark detections. Initially, little is known about these distributions, causing this likely overfitting; however, once several observations of a landmark pair are acquired, the distributions converge to a stable state. With this in mind, we draw on the approach proposed in [6] and complement the observed distribution θ_o with a prior distribution θ_p to estimate the true distribution θ_t:

$$\theta_t = k_n\theta_o + (1 - k_n)\theta_p. \tag{6}$$

Here, k_n is a weighting term which acts to blend the prior and observed distributions, and is a function of n, the number of observations of the landmark pair so far. We assume a Gaussian prior with mean equal to the mean of the observed distribution after n observations. The standard deviation is learned independently for each of the SID and RIA distributions, on a separate training sequence, and set at the 95^{th} percentile of the standard deviations for all tracked landmark pairs in this sequence (obtained as 17.0 and 14.0 for SID and RIA, respectively). The initial value of k_n in Eq. 6 is set to 0.5, and it is updated as $k_n = 0.5 + 0.003n$ (until $k_n = 1$), where blending rate 0.003 between the prior

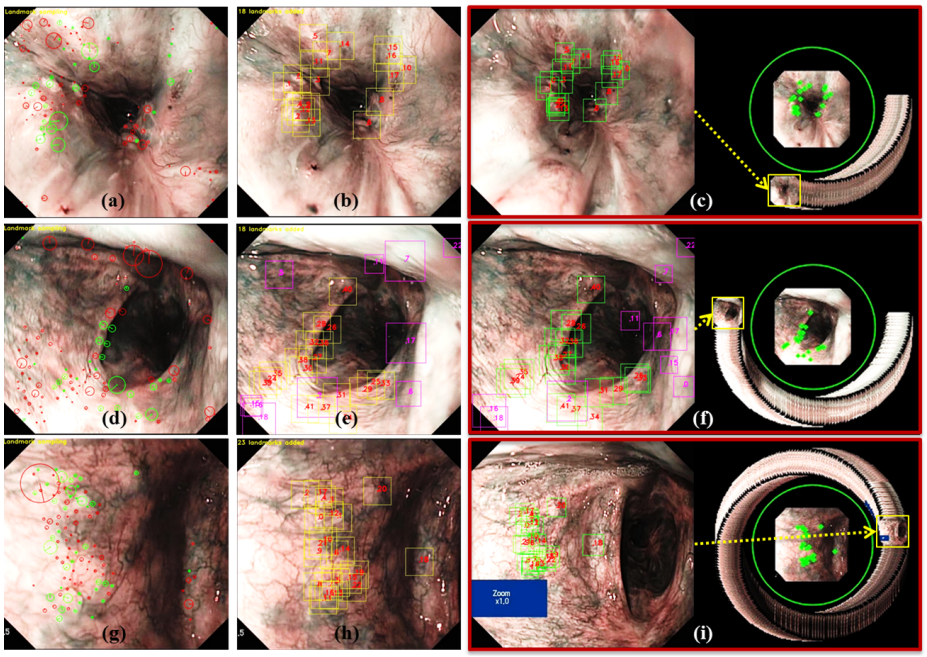

Fig. 3. (a,d,g) Keyframes for landmark sampling. Red features represent unstable features. Green features are the tracked features that are sampled landmarks (yellow squares) in (b,e,h). Purple squares are false positives; (c,f,i) Scene association examples. Green squares are detected landmarks in the query image, and green dots are the corresponding landmarks in the scene. See supplementary videos for details[1].

and observed distributions was obtained by observing the stability of distribution entropies in the separate training sequence as further observations are made.

3 Experiments and Results

Before conducting *in vivo* experiments, the parameters of our approach need to be specified. The initial region sizes of landmarks are defined as 70×70 (in pixels). The bin sizes of SID or RIA distributions are discretised at one pixel and one degree, respectively. The x-axis range of SID distribution is $[0, 440]$ (in pixels), which is obtained by finding the maximum distance between landmarks and the minimum feature size. The x-axis range of RIA distribution is $[0, 359]$ (in degrees). In this paper, the minimum acceptable scores for landmark recognition T_l and scene association T_s are set to 0.5 and 0.25, respectively.

***In vivo* Experiments.** For validation, the proposed approach was tested on four sequences of *in vivo* GI videos. These videos were collected during different

[1] http://www.imperial.ac.uk/hamlyn/surgicalvision

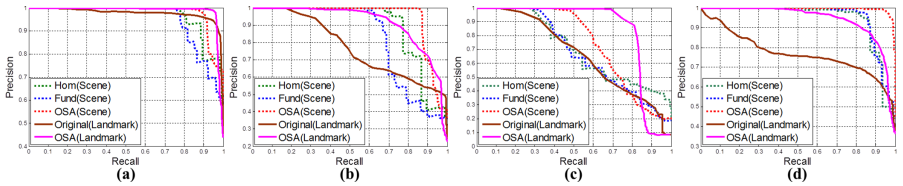

Fig. 4. (a-d) Precision-recall curves of landmark recognition and scene association with four *in vivo* GI videos.

Table 1. Quantitative results of the average precision and maximum recall values at 100% precision.

	Landmark Recognition				**Scene Association**					
	Average Precision		Max Recall		Average Precision			Max Recall		
	OSA	Original	OSA	Original	OSA	Hom	Fund	OSA	Hom	Fund
Video 1	**0.99**	0.98	**0.38**	0.14	**0.98**	0.96	0.95	**0.83**	0.74	0.70
Video 2	**0.91**	0.77	**0.26**	0.17	**0.94**	0.89	0.82	**0.87**	0.68	0.58
Video 3	**0.84**	0.67	**0.21**	0.11	**0.74**	0.70	0.67	**0.40**	0.27	0.30
Video 4	**0.94**	0.77	**0.38**	0.01	**0.99**	0.94	0.95	**0.84**	0.52	0.65

GI procedures using an Olympus Narrow Band Imaging (NBI) endoscope. As shown in Fig. 3, for landmark sampling, features are tracked (within episodes) from the keyframes in Figs. 3 (a,d,g) to the images shown in Figs. 3 (b,e,h). Once the locations of landmarks are found, they are then tracked and learned as square regions. It can be seen from Fig. 3 (e) that our approach is able to reject wrong landmark detections (purple squares) generated from appearance classifiers. After creating a scene from an episode, a query image is then classified and added to the correct scene (Figs. 3 (c,f,i)). As our approach updates the SID and RIA distributions online, it is robust to tissue deformation in Fig. 3 (c) and fast camera motion in Fig. 3 (i). Quantitative comparison and evaluation have also been conducted. The ground truth data of landmark detection and scene association were obtained from expert observers. For landmark detection, we compare the results of the proposed online scene association (OSA) with the orignal [1] (without pairwise geometry learning). For scene association, our approach (OSA) is compared with two standard image matching methods [11]: homography (Hom) and fundamental matrix (Fund). Both methods perform kNN feature matching between images and then use Random Sample Consensus (RANSAC) to find an optimal 3D relationship between the query image and a stored scene image (defined at the end of the scene's episode). Here, the score is the number of RANSAC inliers.

Precision-recall results were generated for recognition of both landmarks and scenes, by ranking landmark or scene matches based on their respective scores (Eqs. 1 and 5). These are shown in Figs. 4 (a-d) (Video 1-4), and the average precision and maximum recall values (at 100% precision) are presented in Table 1. It is evident that our method outperforms all the other methods, with average precisions $[0.84, 0.99]$ for landmark recognition and $[0.74, 0.99]$ for scene association. It should be noted that, the classification overfitting mentioned in Section

2.4 causes the recall values of our landmark detection to be slightly smaller than the original (Fig. 4 (c)), in the precision interval $[0.1, 0.3]$. Nevertheless, in realistic clinical scenarios, the max recall values (at 100% precision) are much more informative, as the clinicians require high levels of confidence ($\sim$100% precision) on the returned scene associations.

4 Conclusion

In this work, we have proposed a scene association approach for endoscopic navigation. Our method samples visual landmarks on the tissue surface using a keyframe-based tracking scheme. An appearance classification scheme has been adopted for long-term landmark detection. To achieve online scene association, the pairwise geometrical relationships between landmarks are learned in a generative model, which is robust to tissue deformation and fast camera motion. Detailed experimental comparison and evaluation have been conducted on *in vivo* GI videos. It has been shown that our approach effectively rejects the wrong detections from appearance classifiers, and simultaneously achieves online scene association, which allows the approach to be performed on-the-fly during endoscopic examinations.

References

1. Ye, M., Giannarou, S., Patel, N., Teare, J., Yang, G.Z.: Pathological site retargeting under tissue deformation using geometrical association and tracking. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, vol. 8150, pp. 67–74. Springer Berlin Heidelberg (2013)
2. Atasoy, S., Mateus, D., Meining, A., Yang, G., Navab, N.: Endoscopic video manifolds for targeted optical biopsy. IEEE Trans. Med. Imag. 31(3), 637–653 (2012)
3. Kwitt, R., Vasconcelos, N., Rasiwasia, N., Uhl, A., Davis, B., Häfner, M., Wrba, F.: Endoscopic image analysis in semantic space. Med. Image Anal. 16(7), 1415–1422 (2012)
4. André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: A smart atlas for endomicroscopy using automated video retrieval. Med. Image Anal. 15(4), 460–476 (2011)
5. Johns, E., Yang, G.Z.: From images to scenes: Compressing an image cluster into a single scene model for place recognition. In: ICCV. pp. 874–881 (2011)
6. Johns, E., Yang, G.Z.: Generative methods for long-term place recognition in dynamic scenes. Int. J. Comput. Vision 106(3), 297–314 (2014)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
8. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-backward error: Automatic detection of tracking failures. In: ICPR. pp. 2756–2759 (2010)
9. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. 34(7), 1409–1422 (2012)
10. Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.: Fast keypoint recognition using random ferns. IEEE Trans. Pattern Anal. Mach. Intell. 32(3), 448–461 (2010)
11. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, second edn. (2004)