Targeted Optical Biopsies for Surveillance Endoscopies

Selen Atasoy[1,2], Diana Mateus[1], Alexander Meining[3], Guang-Zhong Yang[2], and Nassir Navab[1]

[1] Chair for Computer Aided Medical Procedures (CAMP), TU Munich, Germany
{atasoy,mateus,navab}@cs.tum.edu
[2] Hamlyn Centre for Robotic Surgery, Imperial College London, United Kingdom
{catasoy,gzy}@doc.ic.ac.uk
[3] Klinikum Rechts der Isar, TU Munich, Germany
alexander.meining@lrz.tu-muenchen.de

Abstract. Recent introduction of probe-based confocal laser endomicroscopy (pCLE) allowed for the acquisition of *in-vivo* optical biopsies during the endoscopic examination without removing any tissue sample. The non-invasive nature of the optical biopsies makes the re-targeting of previous biopsy sites in surveillance examinations difficult due to the absence of scars or surface landmarks. In this work, we introduce a new method for recognition of optical biopsy scenes of the diagnosis endoscopy during serial surveillance examinations. To this end, together with our clinical partners, we propose a new workflow involving two-run surveillance endoscopies to reduce the ill-posedness of the task. In the first run, the endoscope is guided from the mouth to the z-line (junction from the oesophagus to the stomach). Our method relies on clustering the frames of the diagnosis and the first run surveillance ($\mathcal{S}1$) endoscopy into several scenes and establishing cluster correspondences accross these videos. During the second run surveillance ($\mathcal{S}2$), the scene recognition is performed in *real-time* and *in-vivo* based on the cluster correspondences. Detailed experimental results demonstrate the feasibility of the proposed approach with 89.75% recall and 80.91% precision on 3 patient datasets.

1 Introduction

Oesophageal adenocarcinoma (OAC) is one of the most rapidly increasing cancers in the Western world with a survival rate of less than 20%. The reason of this low survival rate in OAC is largely due to its late diagnosis. To alleviate this problem, patients diagnosed with a precursor of OAC are required to undergo regular surveillance endoscopies where biopsies are taken from suspicious tissue regions. The introduction of the new probe-based confocal laser endomicroscopy (pCLE) enabled real-time visualisation of cellular structures in-vivo. Despite their established advantages, these *optical biopsies* also introduce new challenges into the existing gastro-intestinal (GI) endoscopy workflow. Due to their non-invasive nature, re-targeting the same biopsy locations in subsequent surveillance examination becomes very challenging. Recently, several methods have been proposed

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 83–90, 2011.

for addressing the re-localization problem within one intervention [2,1,11,3]. The application of such localization methods to a *new surveillance* GI endoscopy requires real-time recognition of the frames containing previously targeted biopsy sites. The major challenge of performing scene recognition between the diagnosis and surveillance endoscopies is the variation in visual appearances of the same scene as demonstrated in Fig.1(a),(d). To address this challenge, we propose a *two-run surveillance endoscopy*. In the introduced workflow, prior to the actual surveillance endoscopy, a first-run surveillance (S1) video is acquired in the same examination. This is a commonly performed process in bronchoscopy [9]. To the best of our knowledge, however, this process has not been applied in GI examinations. In this work, we introduce the two run surveillance schema for GI endoscopies, which allows us to provide an applicable solution for re-targeting the optical biopsy sites in surveillance examinations.

The proposed method first creates scene clusters from the diagnosis and $\mathcal{S}1$ endoscopies and then establishes correspondences between these two videos based on expert's supervision. As the structure of the tissue between the $\mathcal{S}1$ and the actual examination performed in the second run surveillance ($\mathcal{S}2$) remains the same, the visual recognition of a scene becomes a solvable task. Once the query scenes, *i.e.* scenes of the diagnosis endoscopy which need to be recognized, are defined, recognition is achieved based on the guided correspondences.

To facilitate the proposed workflow, an endoscopic scene clustering method proposed in [4] is adapted. To this end, we create a manifold representation of the endoscopic videos by taking into account the visual similarities and the temporal relations within the video simultaneously. Scene clustering is performed in the low dimensional space using a mixture model method presented in [7]. The accuracy of the method is validated on 3 different patient datasets, where the patient underwent chemotherapy between the acquisitions.

2 Methods

2.1 Proposed Workflow

In this work, we firstly propose a *two-run surveillance endoscopy*. In the introduced schema, prior to the actual surveillance endoscopy, the endoscope is guided from the mouth to the z-line (junction from the oesophagus to the stomach) without acquiring any optical biopsies. The video of this $\mathcal{S}1$ endoscopy is clustered into different endoscopic scenes and used to acquire scene matching between the diagnosis and surveillance endoscopy. This additional step enables the recognition of the same location despite very large variation in the visual appearances of the scene in different examinations as illustrated in Fig.1(a),(d).

Thus, the proposed workflow involves 3 endoscopic videos: diagnosis endoscopy (Fig.1(a)), where the first optical biopsies have been acquired; $\mathcal{S}1$ (Fig.1(d)) which is performed to provide matches between the endoscopic scene clusters; and the $\mathcal{S}2$ (Fig.1(g)) where the surveillance examination is performed and the previous optical biopsy sites need to be recognized in real-time and *in-vivo*.

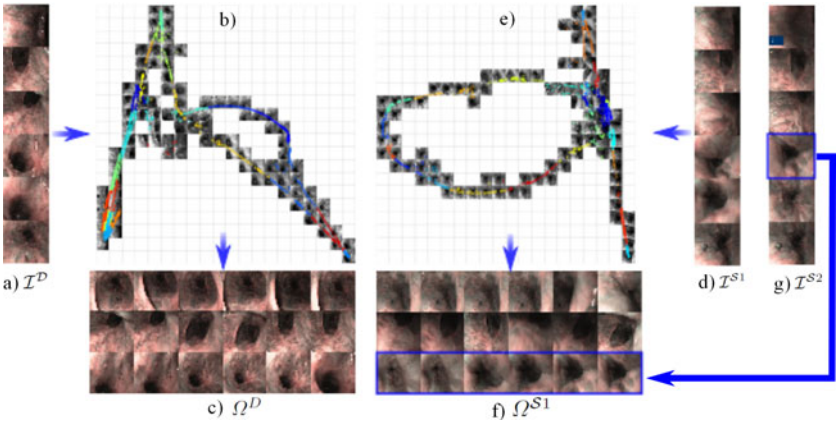


Fig. 1. Proposed workflow. a) Frames from the diagnosis endoscopy. b) 1. and 2. dimensions of the manifold of the diagnosis endoscopy created using vtLPP. Frames showing similar locations are clustered together, where clusters are illustrated with different colors. c) Example clusters of the diagnosis endoscopy, where rows correspond to different clusters. Note that frames of the same scene with different endoscope viewpoint are clustered together whereas different scenes are clustered separately. d) Corresponding scenes of a) in the $\mathcal{S}1$. Rows in a) correspond to rows in d). e) 1. and 2. dimensions of $\mathcal{S}1$ manifold and the computed clusters. f) Frames from the corresponding clusters of c) in the $\mathcal{S}1$. The rows in c) correspond to rows in f). g) Example frames from the $\mathcal{S}2$.

The proposed workflow consists of the following main steps:

1. Clustering of the diagnosis endoscopy into different scenes (Fig.1(a)-(c)),
2. Acquisition of the $\mathcal{S}1$ endoscopy (Fig.1(d)),
3. Clustering of the $\mathcal{S}1$ endoscopy into different scenes (Fig.1(d)-(f)),
4. Selection of the query clusters in the diagnosis endoscopy and their correspondences in the $\mathcal{S}1$ by the endoscopic expert,
5. Nearest neighbour matching and $\mathcal{S}1$ cluster assignment to each frame of the $\mathcal{S}2$ endoscopy in real-time (Fig.1(g)),
6. Notification of the expert during the $\mathcal{S}2$ endoscopy if a frame is assigned to one of the query clusters.

Given the frames of the diagnosis (Fig.1(a)) and of the $\mathcal{S}1$ (Fig. 1(d)) endoscopies, our method first computes a low dimensional manifold representation for each video by taking into account the visual similarities and the temporal relations between the frames. This allows for efficient clustering of the endoscopic scenes. Fig.1(b) and (e) show the 1. and 2. dimensions of the manifolds computed from the diagnosis and $\mathcal{S}1$ endoscopies respectively, where the clusters are illustrated by different colors. Clustering of the frames into different scenes is performed on this manifold representation using a mixture model and the expectation maximization method proposed in [7]. Fig.1(c) shows example clusters from the diagnosis endoscopy where the corresponding clusters in the $\mathcal{S}1$ are illustrated in Fig.1(f). Note the severe change in the appearance of the

scenes between the two examinations. Based on the previously defined diagnosis endoscopy clusters and their correspondences in the $\mathcal{S}1$, the proposed workflow allows for *real-time* and *in-vivo* recognition of the query scenes during the $\mathcal{S}2$.

2.2 Data Representation

Clustering of endoscopic frames using the original image representation is not practical due to the high dimensionality of the data. In [4], the authors propose to recover the underlying non-linear manifold structure of an endoscopic video and to perform the clustering on this low dimensional space. In this work, we approximate the manifold underlying an endoscopic video using the locality preserving projections (LPP) method [10]. In contrast to [4], we define the relations between the frames by taking into account their visual similarities and temporal relations simultaneously and use a probabilistic clustering presented in [7].

LPP first defines an adjacency graph A that captures the pairwise relations $A(i,j)$ between the frames $\mathcal{I}_i$ and $\mathcal{I}_j$, $(i,j \in \{1,\cdots,n\}$, n being the number of data points), and then estimates a mapping to embed the graph into a low dimensional space. In order to simultaneously capture the visual and the temporal relations between the data points, we propose to define the adjacency graph as:

$$A(i,j) = \begin{cases} 1 & \text{if } i \in \mathcal{N}_j^{\text{sim}} \text{ or } i \in \mathcal{N}_j^{\text{temp}} \\ 0 & \text{otherwise} \quad , \end{cases} \tag{1}$$

where $\mathcal{N}_j^{\text{sim}}$ is the k-NN of the j-th data point based on the visual similarities and $\mathcal{N}_j^{\text{temp}}$ states the k-NN based on the temporal order of the frames within the endoscopic video. In this work, we determine the visual similarities using the Euclidean distance and choose $k = 20$ considering the observed endoscope motion. Imposing the proposed temporal constraint assures that frames showing the same scene from different endoscope viewpoints are closely localized on the manifold, even in cases where visual similarities fail to capture their relations. On the other hand, using the visual similarities includes the neighborhood of similar but temporally distant frames, which is reflected in the closed loops on the manifold representations (Fig.1(b),(e)).

Given the adjacency matrix A and the (vectorized) endoscopic frames $\mathcal{I} = [\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_n]$, we approximate the underlying manifold of the endoscopic data using the LPP method [10]. In LPP, first a function basis $w = [w_1, \cdots, w_m]$ is computed based on locally linear approximations of the Laplace-Beltrami operator applied on the dataset by solving the following eigenvalue problem:

$$\mathcal{I}L\mathcal{I}^\top w = \lambda \mathcal{I}D\mathcal{I}^\top w \quad , \tag{2}$$

where D is the diagonal degree matrix with $D(i,i) = \sum_j A(j,i)$ and $L = D - A$ is the graph Laplacian matrix [10]. Then the m dimensional representation $\nu = [\nu_1(i), \cdots, \nu_m(i)]^\top$ of a frame $\mathcal{I}_i$ is estimated by projecting it onto the estimated basis $\nu = w^\top \mathcal{I}_i$. Thus, this method provides an approximation for the Laplacian Eigenmaps (LE) method [5] while it also allows for projection of new data points

onto the manifold. Fig.1(b),(e) illustrate a 2D representation of two endoscopic videos. In the rest of the paper, we refer to our representation as visual and temporal LPP (vtLPP).

2.3 Endoscopic Scene Clustering

Once the low dimensional representations of endoscopic frames are computed, we use the finite mixture models (FMM) method proposed in [7] to compute the clusters. Using FMM, we estimate the probability $P[c(\nu(i)) = C_j]$ of each point $\nu(i)$ belonging to a mixture model (cluster) C_j and assign the cluster with the highest probability $c(\nu(i)) = \arg\max_{C_j} P[c(\nu(i)) = C_j)]$. FMM [7] offers the advantage of automatically detecting the number of clusters. Additionally, FMM models clusters with anisotropic Gaussians, which overcomes the isotropic distribution assumption imposed in clustering algorithms such as K-means [8] and results in elongated clusters. Such clusters efficiently group frames showing the same scene with different viewpoints as shown in Fig.1(b),(c),(e) and (f)).

2.4 Endoscopic Scene Recognition

After computing the clusters of the diagnosis endoscopy $\Omega^{\mathcal{D}} = \{C_1^{\mathcal{D}}, \cdots, C_\alpha^{\mathcal{D}}\}$ and then the ones of the $\mathcal{S}1$ endoscopy $\Omega^{\mathcal{S}1} = \{C_1^{\mathcal{S}1}, \cdots, C_\beta^{\mathcal{S}1}\}$, both clusterings are provided to the endoscopic expert. The set of Q clusters, where an automatic recognition is needed, *i.e.* the query clusters $\{C_q^{\mathcal{D}}\}_{q=1}^{Q} \in \Omega^{\mathcal{D}}$, as well as their correspondences in the $\mathcal{S}1$ endoscopy, $\{C_{\gamma(q)}^{\mathcal{S}1}\} \in \Omega^{\mathcal{S}1}$ (where γ denotes the correspondence relation) are selected by the endoscopic expert.

During the $\mathcal{S}2$, first the image closest to a frame $\mathcal{I}_i^{\mathcal{S}2}$, that is $\mathcal{I}_j^{\mathcal{S}1} = NN(\mathcal{I}_i^{\mathcal{S}2})$, is found by a simple NN matching using Euclidean distances. Then each frame $\mathcal{I}_i^{\mathcal{S}2}$ is assigned the cluster of its NN $c^{\mathcal{S}1}(\mathcal{I}_i^{\mathcal{S}2}) = c^{\mathcal{S}1}(\mathcal{I}_j^{\mathcal{S}1})$ and, by transition, the corresponding diagnosis endoscopy cluster $c^{\mathcal{D}}(\mathcal{I}_i^{\mathcal{S}2}) = c^{\mathcal{D}}(\mathcal{I}_j^{\mathcal{S}1})$. If a frame is determined to belong to a query cluster $c^{\mathcal{D}}(\mathcal{I}_i^{\mathcal{S}2}) \in \{C_q^{\mathcal{D}}\}$, the expert is notified and all frames of the corresponding diagnosis endoscopy cluster $\{\mathcal{I}_k^{\mathcal{D}} | c^{\mathcal{D}}(\mathcal{I}_k^{\mathcal{D}})\}$ are retrieved. This proposed workflow thus allows for including the expert's supervision in defining the query scenes and their correspondences in the $\mathcal{S}1$ without involving any training. This is an important property, since long training processes would not be feasible for routine clinical applications.

3 Experiments and Results

Experiments were performed on 3 narrow-band imaging (NBI) patient datasets acquired at 3 different examinations of the same patient. The patient underwent chemotherapy between the examinations, leading to significant changes in the appearance of the tissue as illustrated in Fig.1. Uninformative frames are labeled using the method in [4] and the remaining informative frames (1198, 1833 and 712 frames in 1., 2. and 3. datasets, respectively) are used for the experiments.

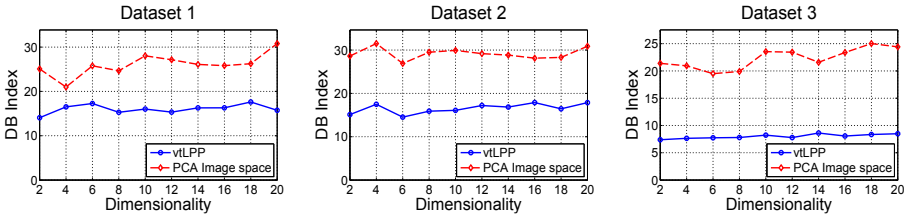


Fig. 2. Evaluation of scene clustering on the proposed representation as compared to the low dimensional image space representation

3.1 Evaluation of Scene Clustering

In order to assess the quality of the clustering, we evaluate the Davis-Bouldin (DB) index [6] which is a commonly used evaluation criteria for clustering algorithms. DB-index measures the relation of the between cluster distances (separability) and within cluster distances (compactness) and is independent of the number of clusters. Smaller DB-indices indicate more compact and separable clusters and are desired. We compare the DB-index of the clustering performed in our vtLPP representation to the one in the PCA representation of the data. Due to its numerical instability, the FMM algorithm [7] is not applicable to very high dimensional data, such as in the original image representation. Therefore, we apply a principal component analysis (PCA) and reduce the dimensionality of the dataset prior to clustering. Using the FMM clustering in [7], we observed that higher dimensional representations result in less number of clusters. Therefore, the evaluation of the DB-index is performed by varying the dimensionality from 2 to 20 for the two methods. Fig.2 shows that for all number of dimensions and for all datasets, the proposed representation results in significantly smaller DB-indices indicating more compact and better separated clusters.

3.2 Evaluation of Scene Recognition

For quantitative analysis we perform 3 experiments. In each experiment, 40 frames from the surveillance endoscopic video are selected by regularly sampling the frames over time and are used as test frames simulating the $\mathcal{S}2$ endoscopic frames leading to a total recognition of 120 frames. Remaining parts of the surveillance video are defined to be the $\mathcal{S}1$ endoscopy. The results are compared to k-NN matching based on Euclidean distances performed between the $\mathcal{S}2$ and diagnosis endoscopy frames directly, where k is chosen to be equal to the number of frames retrieved by our method. We also performed the NN matching using the normalized cross correlation and did not observe a significant improvement in the recognition results. The true positives (tp) and false positives (fp) are determined by expert visual inspection of the retrieved frames. The false negatives (fn) of each method is defined relatively, as the number of frames that one method is able to correctly retrieve but not the other. Recall ($tp/(tp+fn)$) and precision ($tp/(tp+fp)$) values are evaluated for each test frame and mean and standard deviation achieved by both methods is presented in Fig.3. Application

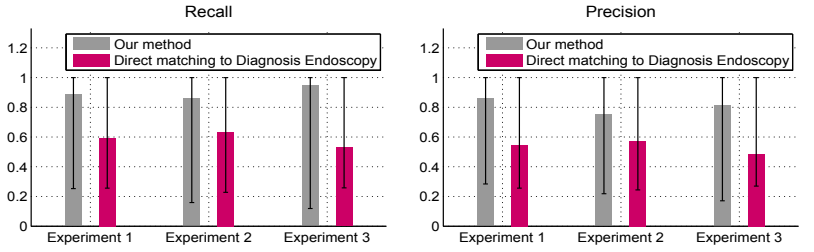
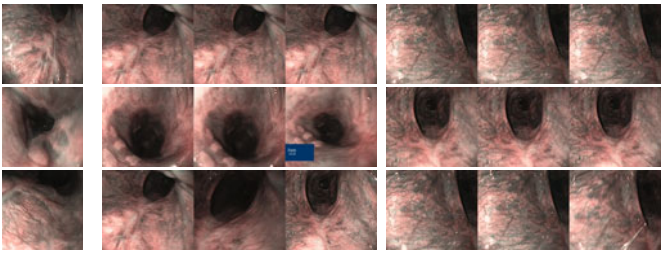


Fig. 3. Mean and standard deviation of recall and precision of the proposed method and of the direct application of the k-NN matching to the diagnosis endoscopy



(a) (b) (c)

Fig. 4. a) Test frames used as $\mathcal{S}2$ endoscopy. b) Recognized frames using our method. c) 3 NN in the diagnosis endoscopy. The rows show corresponding frames in a), b), c).

of the k-NN matching directly between the test frames and the diagnosis endoscopy results in only 58.54% mean recall and 53.58% mean precision. Our proposed method leads to a 89.75% recall and 80.91% precision in average using the same NN matching between the test frames and the $\mathcal{S}1$ endoscopic frames and then applying the cluster correspondences. Examples of the correctly recognized frames using the proposed method in comparison to the direct application of k-NN matching between the $\mathcal{S}2$ and diagnosis videos are demonstrated in Fig.4. Due to the use of our vtLPP representation, the formed endoscopic clusters contain frames showing the same location from different viewpoints and from different parts of the video. This is also reflected in the high recall and precision values of the proposed method.

4 Conclusions

In this work, we present an endoscopic scene recognition method based on two run surveillance endoscopies and scene clustering. The key contributions of this work are two-fold. Technically, we have presented a scene clustering method for endoscopic videos by taking into account both visual similarities and temporal relations in a low dimensional space. Clinically, we have proposed a solution to the challenging problem of re-targeting the optical biopsy sites in *surveillance*

endoscopies. The introduced workflow allows us to create a link between the scenes of the diagnosis and surveillance examinations. This reformulation reduces the very challenging inter-examination re-targeting into the plausible problem of intra-examination frame recognition. The experiments on 3 different patient datasets demonstrate the feasibility of our method to recognize the optical biopsy scenes in surveillance endoscopies.

Acknowledgements. This research was supported by the Graduate School of Information Science in Health (GSISH) and the TUM Graduate School. The authors would like to thank Alessio Dore for valuable discussions.

References

1. Allain, B., Hu, M., Lovat, L., Cook, R., Ourselin, S., Hawkes, D.: Biopsy Site Re-localisation Based on the Computation of Epipolar Lines from Two Previous Endoscopic Images. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 491–498. Springer, Heidelberg (2009)
2. Allain, B., Hu, M., Lovat, L., Cook, R., Vercauteren, T., Ourselin, S., Hawkes, D.: A System for Biopsy Site Re-targeting with Uncertainty in Gastroenterology and Oropharyngeal Examinations. In: Jiang, T., Navab, N., Pluim, J., Viergever, M.A. (eds.) MICCAI 2010, Part II. LNCS, vol. 6362, pp. 514–521. Springer, Heidelberg (2010)
3. Atasoy, S., Glocker, B., Giannarou, S., Mateus, D., Meining, A., Yang, G.Z., Navab, N.: Probabilistic Region Matching in Narrow-Band Endoscopy for Targeted Optical Biopsy. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 499–506. Springer, Heidelberg (2009)
4. Atasoy, S., Mateus, D., Lallemand, J., Meining, A., Yang, G.Z., Navab, N.: Endoscopic Video Manifolds. In: Jiang, T., Navab, N., Pluim, J., Viergever, M.A. (eds.) MICCAI 2010, Part II. LNCS, vol. 6362, pp. 437–445. Springer, Heidelberg (2010)
5. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Comput. 15(6), 1373–1396 (2003)
6. Davies, D., Bouldin, D.: A Cluster Separation Measure. IEEE Trans. on Pattern Anal. (2), 224–227 (1979)
7. Figueiredo, M., Jain, A.: Unsupervised learning of finite mixture models. IEEE Trans. on Pattern Anal. 24(3), 381–396 (2002)
8. Hartigan, J., Wong, M.: A k-means clustering algorithm. Journal of the Royal Statistical Society C 28(1), 100–108 (1979)
9. Häussinger, K., Ballin, A., Becker, H., Bölcskei, P., Dierkesmann, R., Dittrich, I., Frank, W., Freitag, L., Gottschall, R., Guschall, W., Hartmann, W., Hauck, R., Herth, F., Kirsten, D., Kohlhäufl, M., Kreuzer, A., Loddenkemper, R., Macha, N., Markus, A., Stanzel, F., Steffen, H., Wagner, M.: Recommendations for Quality Standards in Bronchoscopy. Pneumologie 58(5), 344 (2004)
10. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.: Face Recognition using Laplacianfaces. IEEE Trans. on Pattern Anal. 27(3), 328–340 (2005)
11. Mountney, P., Giannarou, S., Elson, D., Yang, G.-Z.: Optical Biopsy Mapping for Minimally Invasive Cancer Screening. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5761, pp. 483–490. Springer, Heidelberg (2009)