Endoscopic Video Manifolds for Targeted Optical Biopsy

Selen Atasoy*, Diana Mateus, Alexander Meining, Guang-Zhong Yang, and Nassir Navab

***Abstract*—Gastro-intestinal (GI) endoscopy is a widely used clinical procedure for screening and surveillance of digestive tract diseases ranging from Barrett's Oesophagus to oesophageal cancer. Current surveillance protocol consists of periodic endoscopic examinations performed in 3–4 month intervals including expert's visual assessment and biopsies taken from suspicious tissue regions. Recent development of a new imaging technology, called probe-based confocal laser endomicroscopy (pCLE), enabled the acquisition of *in vivo* optical biopsies without removing any tissue sample. Besides its several advantages, i.e., noninvasiveness, real-time and *in vivo* feedback, optical biopsies involve a new challenge for the endoscopic expert. Due to their noninvasive nature, optical biopsies do not leave any scar on the tissue and therefore recognition of the previous optical biopsy sites in surveillance endoscopy becomes very challenging. In this work, we introduce a clustering and classification framework to facilitate retargeting previous optical biopsy sites in surveillance upper GI-endoscopies. A new representation of endoscopic videos based on manifold learning, "endoscopic video manifolds" (EVMs), is proposed. The low dimensional EVM representation is adapted to facilitate two different clustering tasks; i.e., clustering of informative frames and patient specific endoscopic segments, only by changing the similarity measure. Each step of the proposed framework is validated on three *in vivo* patient datasets containing 1834, 3445, and 1546 frames, corresponding to endoscopic videos of 73.36, 137.80, and 61.84 s, respectively. Improvements achieved by the introduced EVM representation are demonstrated by quantitative analysis in comparison to the original image representation and principal component analysis. Final experiments evaluating the complete framework demonstrate the feasibility of the proposed method as a promising step for assisting the endoscopic expert in retargeting the optical biopsy sites.**

***Index Terms*—Classification, clustering, gastro-intestinal (GI)-endoscopy, manifold learning, optical biopsy.**

I. Introduction

GASTRO-INTESTINAL (GI) endoscopy is a widely used clinical technique for visualizing the digestive tract. Current diagnosis and surveillance of GI diseases, ranging from Barrett's Oesophagus (BO) to oesophageal or colorectal cancer, are performed by visual assessment via an endoscopic examination followed by necessary biopsies. Patients diagnosed with oesophageal cancer or with BO[1] undergo periodic endoscopic examinations. During these procedures the oesophageal tissue is examined under endoscopic guidance and biopsies are acquired from suspicious regions. In the current procedure, follow-up of the disease is achieved by periodic surveillance endoscopies in 3–4 months intervals. In these surveillance endoscopies, biopsies are acquired from the same sites as in the first screening examination in order to compare the histological diagnosis.

Recently, a new technology called probe-based confocal laser endomicroscopy (pCLE) became available which allows for the *in vivo* visualization of the tissue at a cellular level. A fibered confocal microprobe is inserted through the instrument channel of a standard endoscope (Fig. 1). In contact with the tissue, pCLE visualizes the tissue at a microscopic scale allowing for "optical biopsies." Due to its noninvasive nature, the real-time and *in vivo* feedback, pCLE provides significant advantages over the conventional biopsy. High agreement between the results of optical biopsy and conventional histopathology suggest that pCLE will be increasingly used in daily clinical routine [2], [3].

Besides these advantages, introduction of pCLE into the workflow of endoscopic procedures also induces new challenges. The interpretation of the optical biopsies for *in vivo* diagnosis is a new concept for the endoscopic experts. In [4]–[6], André *et al.* present image and video retrieval methods for pCLE in order to support the endoscopic expert in establishing an *in vivo* diagnosis. A system to facilitate the training for *in vivo* diagnosis based on optical biopsies has also been investigated in [7]. A further challenge introduced by pCLE is the retargeting of previous optical biopsy sites. In contrast to the conventional biopsy, the noninvasive optical biopsies do not leave any scar on the tissue which was being used as landmarks by the endoscopic experts for recognizing previous biopsy sites in surveillance examinations. Several approaches have been proposed for point-based relocalization of the fibered confocal microprobe within an endoscopic frame [8]–[10]. Allain *et al.* present a method for probe relocalization based on the epipolar geometry between endoscopic frames [8], [9]. Mountney *et al.* propose introducing simultaneous localization and mapping (SLAM) framework in order to create a 3-D model of the tissue surface and to provide an augmented view for re-localization of the optical probe [10]. In a previous study, we presented a deformable wide-baseline matching method that provides point correspondences between the current endoscopic view and previous optical biopsy locations [11].

[1]BO refers to an abnormal change in the tissue [1] and is the only recognized precursor of the highly lateral oesophageal cancer.

Manuscript received August 25, 2011; revised October 19, 2011; accepted October 25, 2011. Date of publication November 02, 2011; date of current version March 02, 2012. This research was partially supported by TUM Graduate School and SFB 824. *Asterisk indicates corresponding author.*

*S. Atasoy is currently with the Chair for Computer Aided Medical Procedures, Technische Universität München, 85748 München, Germany, and with the Hamlyn Centre for Robotic Assisted Surgery, Imperial College London, SW7 2AZ London, U.K. (e-mail: atasoy@cs.tum.edu; catasoy@imperial.ac.uk).

D. Mateus and N. Navab are with the Chair for Computer Aided Medical Procedures, Technische Universität München, 85748 München, Germany (e-mail: mateus@cs.tum.edu; navab@cs.tum.edu).

A. Meining is with the Medizinische Klinik und Poliklinik (Gastroenterologie), Klinikum Rechts der Isar, Technische Universität München, 81664 München, Germany (e-mail: alexander.meining@lrz.tum.de).

G.-Z. Yang is with the Hamlyn Centre for Robotic Assisted Surgery, Imperial College London, SW7 2AZ London, U.K. (e-mail: g.z.yang@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2011.2174252

Fig. 1. (a) State-of-the-art flexible endoscope commonly used in upper GI-endoscopic procedures. (b) Confocal microprobe as inserted through the instrument channel of the flexible endoscope. (c) pCLE as viewed from the endoscope while acquiring optical biopsy during upper GI-endoscopy. (d) An example optical biopsy acquired *in vivo* using pCLE during an upper GI-endoscopic procedure.

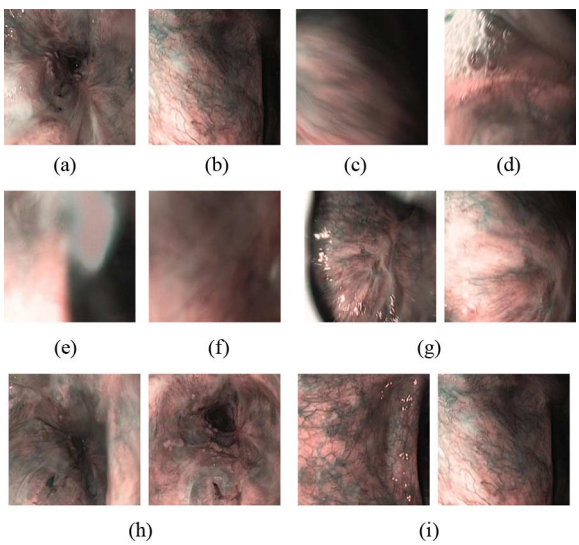

Fig. 2. Several challenges encountered in endoscopic videos (a) and (b) show two examples for informative frames acquired in an upper GI-endoscopy. (c) Motion blur. (d) Bubbles caused by the liquid inside the organ. (e) Specular highlights. (f) Blur caused by out-of-focus. (g)–(i) Two corresponding frames, one from the diagnostic and one from the surveillance endoscopy, showing the same scene for three different PSESs.

In this work, we introduce an endoscopic clustering-classification framework to assist the expert in retargeting the optical biopsy sites. Complementary to the point-based re-localization methods [8]–[11], our framework aims at identifying the frames of a surveillance endoscopy with matching scenes from the previously performed diagnostic endoscopy.

Particular conditions of upper GI-endoscopic videos introduce several challenges into the clustering and classification tasks. Firstly, endoscopic videos suffer from a large number of uninformative frames such as blurry frames due to fast motion or out of focus imaging of the endoscope, frames filled with bubbles caused by the turbuid fluid inside the oesophagus and frames with large specular reflexions [Fig. 2(c)–(f)]. Presence of these uninformative frames leads to poor classification performance. In the literature, detection of uninformative frames has been studied for several endoscopic procedures such as capsule endoscopy [12]–[14], colonoscopy [15], [16]. Main focus of these studies is on defining specific features such as color or texture in order to detect uninformative frames within an endoscopic video. We address this challenge, by clustering the informative and uninformative frames of upper GI-endoscopic videos and allowing the expert to easily choose the frames (clusters) for further processing [Fig. 3(a)–(d)].

In endoscopic image classification, the focus is mainly directed towards computer aided diagnosis of polyps [17]–[19] and tumors [20] or detection of endoscopic lesions [21]–[23]. Recently, video summarization using representative frame extraction has also been investigated for wireless capsule endoscopy [14], [24].

In our work, we address the clustering of the informative frames into different patient specific endoscopic segments (PSESs) and then the identification of a new endoscopic frame with one of the PSESs via classification. A PSES is a group of informative frames showing the same scene (part of the oesophagus) in the diagnostic endoscopy. Each PSES is represented by the set of all frames belonging to one cluster and no particular representative frame is chosen as performed in endoscopic video summary approaches. In this way, each PSES contains the frames showing the same segment of the oesophagus from different viewpoints of the endoscope.

The second difficulty in these clustering and classification tasks lies in the large visual variability within the same endoscopic segment and relatively small visual differences between different segments in upper GI-endoscopic videos as illustrated in [Fig. 2(g)–(i)]. To address this problem, we introduce a novel representation of endoscopic videos. This new representation, called "endoscopic video manifolds" (EVMs) and explored previously in our study [25], is created by learning the low dimensional manifold of an endoscopic video.

Our proposed framework involves offline (postprocedural) processing of the endoscopic video acquired during the diagnostic endoscopy in order to define PSESs and online (intra-procedural) classification of new frames acquired during the surveillance endoscopy as belonging to one of these predefined segments. As the first clustering step of the offline processing, frames of the diagnostic endoscopic video are clustered into informative and uninformative clusters on a suitably designed EVM [Fig. 3(a) and (b)]. To this end, a new similarity measure based on the power spectra of the frames is introduced which emphasizes the difference between informative and uninformative frames. Then, labelling of the informative clusters is performed by the endoscopic expert [Fig. 3(c) and (d)]. In the second postprocedural step, the PSESs are defined by clustering frames of the informative clusters only in a new EVM representation [Fig. 3(e) and (f)]. Thus, the definition of PSESs is based on two different clustering steps; i.e., first, clustering of informative/uninformative frames, and second, clustering of different endoscopic scenes. Each of these two clustering steps is performed in an unsupervised manner. However, between the two clustering steps, it is the endoscopic expert who determines the informative and uninformative clusters [Fig. 3(c)]. Finally, a new surveillance endoscopic frame is projected into the low dimensional space and classified as belonging to one of the PSESs [Fig. 3(g)–(i)].

The technical contribution of this work is twofold. Firstly, we introduce a new representation of endoscopic videos, namely the EVMs, which facilitates the clustering and classification tasks. For each clustering task; i.e., clustering of uninformative frames and clustering of endoscopic segments, a suitably designed EVM is created. Due to the particular scheme of the used manifold learning framework, the manifold structure is easily adapted to the addressed task by defining different similarity

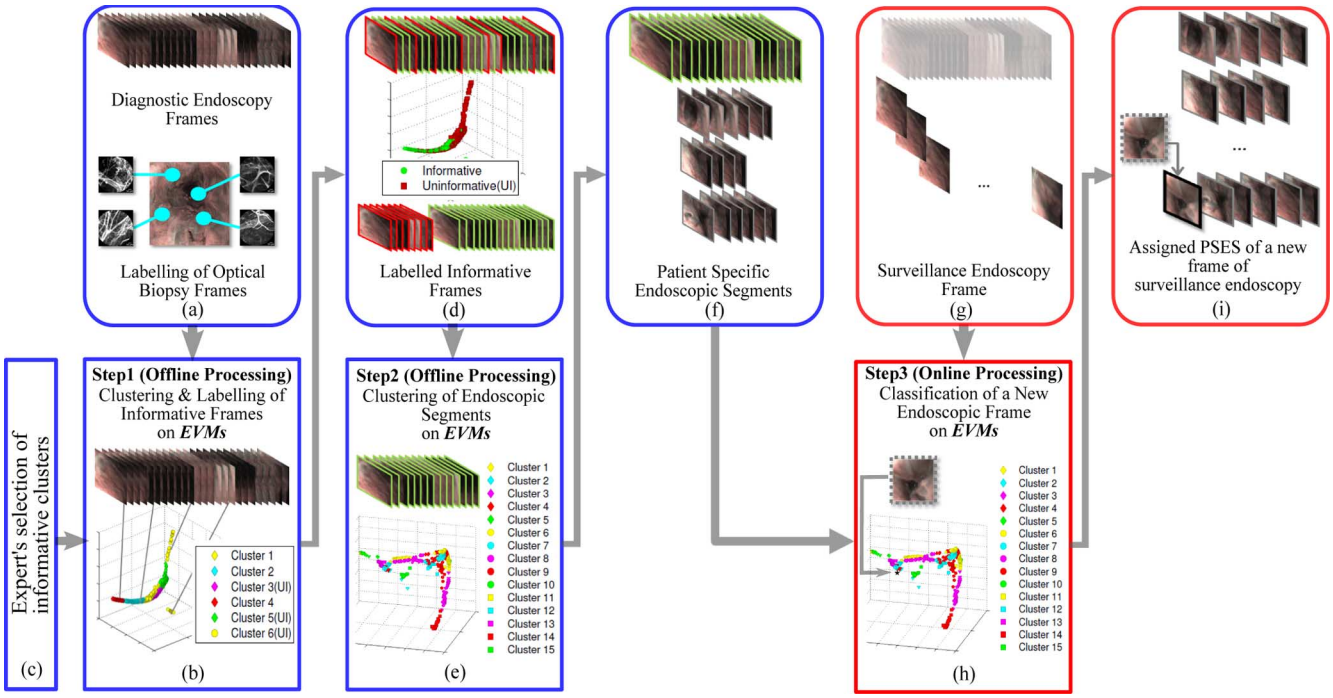

Fig. 3. Proposed clustering-classification framework. (a) Input of the offline processing step consists of the frames of the diagnostic endoscopy and labelling of the frames where an optical biopsy has been acquired. (b) Clustering of the informative and uninformative frames on the corresponding EVM. (c) Selection of the informative clusters by endoscopic expert. (d) Outcome of the first step of the proposed framework; labelling of the informative and uninformative frames of the diagnostic endoscopy. (e) Clustering of the informative frames on a suitable designed EVM. (f) PSESs corresponding to clusters of informative frames. Together with previously labelled uninformative clusters, PSESs form the classes to assign and are first input of the online processing stage. (g) Frames of the surveillance endoscopic video being the second input of the online processing stage. (h) Classification of a new endoscopic frame using a nearest neighbor classifier in the low dimensional space. (i) Outcome of the last step of the framework; assigned PSES of the new endoscopic frame.

measures between data points. Secondly, we introduce two similarity measures into the manifold learning framework; i.e., energy histogram similarity S_{EH} leading to an EVM representation that allows for clustering uninformative frames of an endoscopic video, and normalized cross correlation S_{NCC} resulting in an EVM where different endoscopic scenes are efficiently clustered to define the PSESs. Effectiveness of the introduced representation is demonstrated with quantitative evaluation performed for each step as well as for the overall framework. In our experiments, we demonstrate that the created EVM representation not only leads to a dimensionality reduction of the endoscopic data but also yields more accurate clustering and classification results compared to the original image representation. From the medical point of view, we propose a clustering-classification framework for assisting the endoscopic expert in retargeting previous optical biopsy sites during surveillance upper GI-endoscopic procedures.

The rest of the paper is organized as follows. In Section II, we provide a short introduction to the theoretical background on manifold learning. Section III presents an overview of the proposed manifold learning framework, whereas its application for each of the addressed tasks are explained in Sections IV– VI. The experiments for each part and for the overall framework are demonstrated in Section VII and a final discussion and conclusion are presented in Section VIII.

II. Theoretical Background

Several medical datasets with smooth variation between individual data points, for instance the frames of a upper GI-endoscopic video, do not span the entire high dimensional image space but lie on or near a lower dimensional manifold. Nonlinear manifold learning methods seek to find a low dimensional representation of such datasets while preserving their local structure. Since the early development of nonlinear dimensionality reduction [26]–[29], manifold learning methods have been successfully applied to several medical applications [25], [30]–[42]. A comprehensive review of existing manifold learning techniques can be found in [43]–[45]. These nonlinear manifold learning methods approximate the low dimensional manifold, which the data lies on, using a graph structure and compute the mapping of each data point from the high dimensional input space to a low dimensional space. Different techniques in the literature, differ in their approach for computing this nonlinear mapping from the high to the low dimensional spaces. Spectral methods such as the Laplacian Eigenmaps applied to compute the EVMs, approximate the manifold with a neighborhood graph created by connecting each data point with its k-nearest neighbors according to some similarity measure. The graph can be represented by a matrix and the mapping from the high to low dimensional space is obtained based on the eigenvectors of this matrix. In fact, changing the pairwise similarities between the data points, also changes the neighborhood graph and thus the structure of the approximated manifold. Therefore, by introducing two new similarity measures; i.e., energy histogram similarity S_{EH} and normalized cross correlation S_{NCC}, we adapt the structure of the manifold to each of the two addressed clustering tasks.

Let the pairwise similarities be represented with a matrix S, where each element $S(i, j)$ denotes the measured similarity be-

tween the ith and jth data point. If the similarities are defined by a metric measure, they give rise to a symmetric positive definite matrix S. In this study, we explore different metric measures for defining the structure of the EVMs.

III. Creating Endoscopic Video Manifolds

Image classification and retrieval for large databases has been intensively studied for computer vision applications (examples are [46]–[49]). The set of images in these large databases spans a very high dimensional image space. If an image is considered as a data point in the high dimensional input space $I \in \mathbb{R}^{(w \times h)}$, where w and h are the width and height of the image, respectively, then the set of all possible images spans a $(w \times h)$ dimensional space. In our case, however, due to the temporal continuity of the endoscopic video, and therefore the large similarity between frames showing the same location, the frames $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\} \in \mathbb{R}^{(w \times h)}$ of an endoscopic video $\mathcal{I}$ do not span this high dimensional space completely but lie on or near a low dimensional manifold. Thus, the intrinsic dimensionality of the endoscopic data is much smaller than the number of degrees-of-freedom (DoF) in the original high dimensional representation $(w \times h)$.

To recover the corresponding low dimensional representation, nonlinear manifold learning methods such as [26]–[29] can be used. These methods rely on a nonlinear map to embed the high dimensional data into a low dimensional space. In the presence of complex nonlinear relations between data points (in our framework between endoscopic frames), these methods provide an important tool for dimensionality reduction [44], [45]. In this work, we compute this low dimensional EVM with the following steps.

Step 1) Defining the similarities between the data points.
Step 2) Constructing the adjacency graph.
Step 3) (Optional): Including temporal constraints.
Step 4) Computing the Eigenmaps.
Step 5) Mapping the data to the low dimensional EVM representation.
Step 6) Projecting new data points onto the low dimensional representation.

Steps 2, 4, and 5 formulate the Laplacian Eigenmaps method as proposed by Belkin and Niyogi [28]. We introduce the use of two different similarity measures in Step 1 into this framework in order to create well-structured manifolds permitting the clustering informative frames and different endoscopic segments. Furthermore, we provide an optional step (Step 3) that allows for including the temporal constraints.

A. Defining the Similarities

For each pair $(\mathcal{I}_i, \mathcal{I}_j)$, of the given n data points $i, j \in \{1, \ldots, n\}$, first a similarity measure is defined $\mathcal{S} : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}$. $\mathcal{S}$ determines which images are considered to be similar and therefore kept as neighbors on the manifold. The choice of the similarity measure determines the structure of the manifold and should be designed carefully for each particular application. In the Sections IV-A and V-A we present the similarity measures designed for the addressed clustering tasks.

B. Constructing the Adjacency Matrix

Given the similarity matrix S, where the values $S(i, j)$ stand for pairwise similarities between the frames $\mathcal{I}_i$ and $\mathcal{I}_j$ according to the chosen similarity measure $\mathcal{S}(\mathcal{I}_i, \mathcal{I}_j)$, first, the k-nearest neighbors of each data point are computed. Then, the adjacency matrix is created as

$$W(i,j) = \begin{cases} 1, & \text{if } \mathcal{I}_i \in \mathcal{N}_j^{\text{sim}} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\mathcal{N}_j^{\text{sim}}$ is the set of k-nearest neighbors of the frame $\mathcal{I}_j$ based on the similarity matrix S.

In [28], the authors propose an optional weighting of the edges of an adjacency graph W with a heat (Gaussian) kernel for a chosen time parameter t. The combinatorial adjacency matrix defined as in (1) corresponds to using a Gaussian kernel weighting with $t = \infty$ [28]. A manifold created with the combinatorial adjacency matrix respects only the presence/absence of a connection (edge) between two data points (nodes) and does not take into account the strength of the connection (the measure of similarity). This is a desired property in the case of EVMs in order to achieve some degree of invariance to camera viewpoint change. For two endoscopic frames acquired with slightly different camera viewpoints, the measured similarity will be lower than their consecutive frames in the video (however still higher compared to an unrelated scene). The combinatorial adjacency graph ensures that the strength of the connection between these frames will be the same as of the to their consecutive frames. Thus, on the EVM, frames with slightly different endoscope viewpoints will also be closely localized and clustered into the same PSES.

The number of nearest neighbors k affects the manifold structure by defining the neighborhood $\mathcal{N}_j^{\text{sim}}$ of a frame $\mathcal{I}_j$. This parameter regulates the connectedness of the adjacency graph and is in general estimated empirically. In our experiments (Section VII-A) we provide a performance evaluation of the clustering for different values of the parameter k.

C. Including Temporal Constraints

In an endoscopic video, there also exist temporal relations between frames. Due to the continuity of the endoscopic video, temporally related frames can be assumed to belong to the same scene. In our previous study, we investigated the enforcement of temporal relations using a similarity measure based on the optical flow field between two frames [25]. This measure is based on the smoothness of the optical flow field between two temporally related frames and leads to the clustering of sequential frames together as long as the endoscope motion between them is a smooth vector field. This temporal constraint, however, does not allow for clustering temporally distant frames into the same cluster and therefore is not suitable for defining the PSESs. In this work, we present a method for combining the temporal constraints with the visual similarities simply by defining an additional neighborhood $\mathcal{N}_j^{\text{temp}}$ based on the temporal order of the frames within the endoscopic video. In order to account for

both, the temporal relations as well as visual similarities, the adjacency matrix is defined as

$$W(i,j) = \begin{cases} 1, & \text{if } \left((\mathcal{I}_i \in \mathcal{N}_j^{\text{sim}}) \text{ or } (\mathcal{I}_i \in \mathcal{N}_j^{\text{temp}})\right) \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

In our experiments, we use the same number k for defining the visual $\mathcal{N}_j^{\text{sim}}$ and the temporal $\mathcal{N}_j^{\text{temp}}$ neighborhoods. Enforcing this temporal constraint leads to clustering of temporally close frames 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 and thus allows for grouping frames from different parts of the video together if they lead to high visual similarities. In the rest of the paper we refer to the EVMs with temporal constraints as visual and temporal endoscopic video manifolds (vtEVMs).

D. Computing Laplacian Eigenmaps

After creating the adjacency graph, the graph Laplacian L is computed as

$$L = D - W \quad (3)$$

where D represents the diagonal degree matrix with elements $\forall i,j \in \{1,\ldots,n\}$

$$D(i,i) = \sum_{j=1}^{n} W(i,j). \quad (4)$$

The Laplacian Eigenmaps are determined as the eigenvectors $\{v_1,\ldots,v_d\}$ of the Laplacian matrix L. As each of the Laplacian Eigenmaps $\{v_1,\ldots,v_d\}$, with d being the dimensionality of the manifold, solves the generalized eigenvalue problem $Lv = \lambda Dv$, they minimize the following objective function [28]:

$$\sum_{i,j\in\{1,\ldots,n\}} (v(i) - v(j))^2 W(i,j) \quad (5)$$

where $\mathbf{v}_i = [v_1(i),\ldots,v_d(i)]^\top$ defines the d-dimensional representation of the data point $\mathcal{I}_i$.

The eigenvectors $\{v_1,\ldots,v_d\}$ are the discrete equivalents of the eigenfunctions of the Laplace–Beltrami operator in the continuous domain as applied onto the manifold. Note that using the Laplacian Eigenmaps [28] we compute *the value of the eigenfunction at given data points and not the continuous eigenfunctions themselves*. This distinction and its effects are discussed in detail in the next section (Section III-F).

E. Endoscopic Video Manifold (EVM) Representation

The d-dimensional ($d \ll (w \times h)$) representation of a frame $\mathcal{I}_i$ on the EVM is given by ith entries of the d eigenvectors of the Laplacian matrix L corresponding to the d-smallest eigenvalues

$$\mathbf{v}_i = [v_1(i),\ldots,v_d(i)]^\top. \quad (6)$$

Ideally the dimension d of the new representation should be chosen to be equal to the intrinsic dimensionality of the data; i.e., to the minimum number of parameters needed in order to capture all relevant information about the data. In spectral nonlinear manifold learning methods, the dimensionality of the manifold is generally estimated based on the spectral gap in the eigenvalues of the corresponding eigenvectors. Although this choice is theoretically motivated, for many practical datasets, the spectrum of the Laplacian matrix L exhibits a near to continuous spectrum and does not provide a clear estimation of dimensionality. Therefore, in Sections VII-A1 and VII-B1, we provide an evaluation on the effect of the dimensionality d for clustering uninformative frames and PSES, respectively.

F. Projection of New Data Points

The original high dimensional representation of the data; i.e., in $(w \times h)$ dimensional space does not allow for a robust and fast classification due to the curse of dimensionality [50]. Although the dimensionality can be reduced as described in the previous steps with nonlinear manifold learning, these techniques require all data points to be available to compute their low dimensional representation. Unlike linear methods such as PCA, nonlinear manifold learning techniques compute the projections of the data points onto the low dimensional representation *without explicitly estimating the mapping function to project the data from the high to the low dimensional space*; i.e., the projection of a data point $\mathcal{I}_i$ into the low dimensional space is computed as $\mathbf{v}_i = \nu(\mathcal{I}_i)$, $\mathcal{I}_i \in \mathcal{I}$ without explicitly estimating the continuous nonlinear mapping $\nu : \mathbb{R}^{(w\times h)} \rightarrow \mathbb{R}^d$ from the high dimensional space $\mathbb{R}^{(w\times h)}$ into the low dimensional representation $\mathbb{R}^d$. Therefore, the projection $\mathbf{v}_s$ of a new data point $\mathcal{I}_s \notin \mathcal{I}$ cannot be easily computed by evaluating the mapping function $\nu(\mathcal{I}_s) = \mathbf{v}_s$ for an unknown data point $\mathcal{I}_s$.

The Locality Preserving Projections (LPP) method [50] estimates the optimal linear mapping from the high to low dimensional representations by optimizing the same objective function as defined in (5). However, the objective function is solved for the optimal linear transformation $\widetilde{\nu}$ instead of solving directly for the low dimensional coordinates.

Let the projection $v(i)$ of a high dimensional data point $\mathcal{I}_i$ onto the one dimensional line be defined using a linear mapping

$$v(i) = \widetilde{\nu}(\mathcal{I}_i) = \mathbf{t}^\top \mathcal{I}_i \quad (7)$$

where $\mathbf{t}$ denotes the (linear) transformation vector.

Substituting (7) in the objective function given in (5) results in

$$\sum_{i,j} \left(v(i) - v(j)\right)^2 W(i,j) = \sum_{i,j} \left(\mathbf{t}^\top \mathcal{I}_i - \mathbf{t}^\top \mathcal{I}_j\right)^2 W(i,j) \quad (8)$$

which in turn can be expressed in matrical form

$$\sum_{i,j} (v(i) - v(j))^2 W(i,j) = \mathbf{t}^\top \mathcal{I} L \mathbf{t} \mathcal{I}^\top \quad (9)$$

where $\mathcal{I} = [\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n]$ denotes the complete data matrix. Thus, the transformation vector $\mathbf{t}$ is estimated by the solution of the generalized eigenvalue problem

$$\mathcal{I} L \mathcal{I}^\top \mathbf{t} = \lambda \mathcal{I} D \mathcal{I}^\top \mathbf{t}. \quad (10)$$

For the detailed derivation of (9) and (10) from (8) we refer to [50].

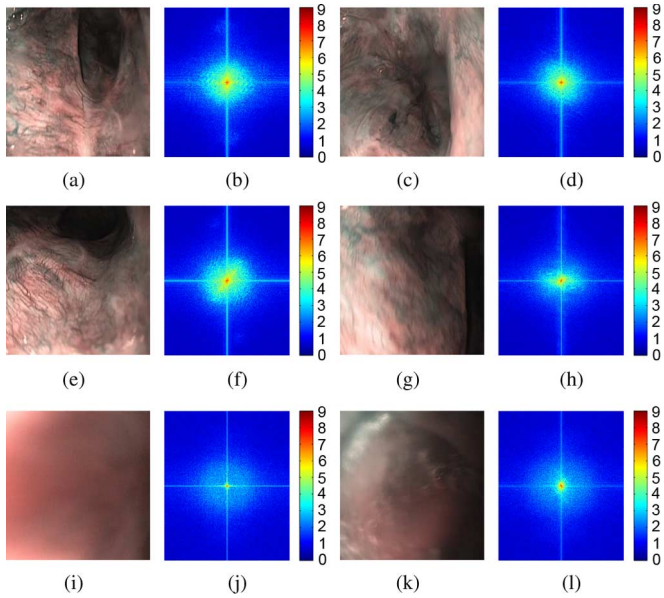

Fig. 4. (a), (c), and (e) Endoscopic frames in ideal conditions acquired by a state-of-the art GI endoscope. (b), (d), and (f) Power spectrum of the ideal frames. (g), (i), and (k) Examples of uninformative endoscopic frames with motion blur, out-of-focus blur, and with bubbles, respectively. (h), (j), and (l) Power spectrum of the ideal frames. For the clarity of the visualization, the dc-components (the part of the spectrum corresponding to a constant wave) are removed and the logarithm of $1+$ magnitude of the spectrum is shown for all power spectrum images.

For the classification step, frames of the diagnostic endoscopy and of the new surveillance endoscopy are projected into the low dimensional representation by applying the linear mapping $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_d]$

$$\tilde{\nu}(\mathcal{I}_s) = \mathbf{T}^{\mathsf{T}} \mathcal{I}_s = \mathbf{v}_s. \tag{11}$$

IV. Clustering and Labelling Uninformative Frames

The first step of our proposed framework consists of clustering uninformative frames on the accordingly created EVM. Given a diagnostic endoscopic video, its EVM representation is computed as explained in Section III. A novel similarity measure (energy histogram similarity) is introduced, which was first explored in our previous study [25]. This similarity measure emphasizes the difference between an informative and uninformative frame and thus leads to better separation of these two different classes on the created EVM.

A. Energy Histogram Similarity

In order to create an EVM, where the uninformative frames are closely localized, the similarity measure used for manifold learning (Section III) needs to be designed such that it yields a low inter-class similarity between informative and uninformative clusters and a high intra-class similarity within each individual informative or uninformative cluster. To this end, we make use of the information captured in the power spectrum of an image.

In the frequency domain, the energy of an informative frame is more distributed over low and high frequencies [Fig. 4(a)–(f)] compared to an uninformative frame whose energy mainly accumulates in low frequencies [Fig. 4(g)–(l)]. To create a measure of similarity between the power spectra, first the Fourier transform of an endosopic frame $\mathcal{I}_i$ is computed and represented in log-polar coordinates $f_i(\omega, \theta)$, where $\omega \in [\omega_0, \ldots, \omega_x, \ldots, \omega_r]$ is the set of considered frequencies and $\theta \in [0, \ldots, 2\pi]$ is the discrete set of orientations. In order to achieve rotation invariance, the 2-D spectrum is integrated over θ resulting in a $r \times 1$ dimensional vector $\mathcal{F}_i$, r being the number of different frequencies of the discrete Fourier transform

$$\mathcal{F}_i(\omega) = \sum_{\theta=0}^{2\pi} f_i(\omega, \theta). \tag{12}$$

To increase the discrimination between informative and uninformative frames, this rotation-invariant spectrum is further discretized into b bins $\{\gamma_1, \ldots, \gamma_\delta, \ldots, \gamma_b\}$

$$\gamma_\delta = \left\{ \omega_x \middle| \frac{\delta - 1}{b} \leq \omega_x \leq \frac{\delta}{b} \right\}, \ \forall \delta \in \{1, \ldots, n\} \tag{13}$$

and the *Engergy Histogram* $\mathbf{h}_i$ of the frame $\mathcal{I}_i$ is defined as the $b \times 1$ dimensional vector

$$\begin{aligned} \mathbf{h}_i &= [h_{\gamma_1}(\mathcal{F}_i), \ldots, h_{\gamma_b}(\mathcal{F}_i)]^{\mathsf{T}} \\ h_{\gamma_\delta}(\mathcal{F}_i) &= \sum_{\omega_x \in \gamma_\delta} \mathcal{F}_i(\omega_x). \end{aligned} \tag{14}$$

In Section VII-A1, we evaluate the clustering accuracy of uninformative frames for several values of b.

Once the rotation-invariant energy histogram of each frame is computed, the similarity between two histograms is measured based on the cosine similarity measure

$$\begin{aligned} S(\mathcal{I}_i, \mathcal{I}_j) &= S_{\text{EH}}(\mathcal{I}_i, \mathcal{I}_j) \\ S_{\text{EH}}(\mathcal{I}_i, \mathcal{I}_j) &= 1 - \left(\frac{\langle \mathbf{h}_i, \mathbf{h}_j \rangle}{\|\mathbf{h}_j\| \cdot \|\mathbf{h}_j\|} \right) \end{aligned} \tag{15}$$

where $\langle \cdot, \cdot \rangle$ is the dot product and $\|\mathbf{h}\|$ denotes the norm of the $b \times 1$ dimensional histogram vector.

Due to the dot product formulation, the cosine similarity relies on the angle between the two histogram vectors in the b-dimensional space. Thus, the similarity between the two vectors is related to the distribution of the values and not to the absolute values of the vectors. This is an important property as it leads to high similarity between two frames whose spectra have different absolute values but similar distributions, such as two informative (or uninformative) frames showing different locations of the oesophagus. As the distribution of the values will be different for an informative $\mathcal{I}_i$ and uninformative frame $\mathcal{I}_j$ as shown in Fig. 5, the energy histogram similarity measure $S_{\text{EH}}(\mathcal{I}_i, \mathcal{I}_j)$ yields a low similarity due to the use of the cosine similarity measure. Thus computing the manifold with the energy histogram similarity measure leads to a representation where the data points are arranged based on their informativeness.

B. Clustering on the EVMs

After evaluating the energy histogram similarity measure S_{EH} on all pairs of frames, EVMs are computed as described in Section III. Then, a K-means clustering [51] is performed in the EVM representation. Cluster centres are initialized randomly

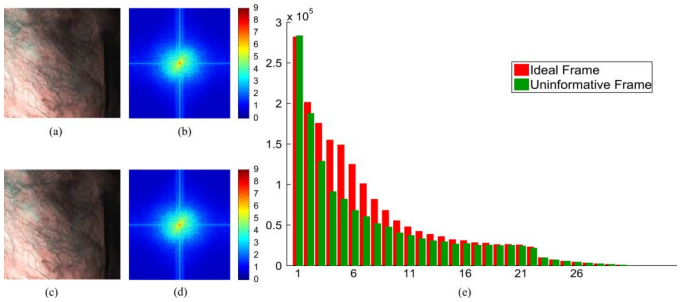

Fig. 5. (a) An informative frame acquired by a state-of-the art GI endoscope and (b) its power spectrum. (c) An endoscopic frame with motion blur. (d) The power spectrum of the motion blurred frame. (e) Rotation-invariant energy histograms of an informative and uninformative frame computed with 30 bins.

and 100 trials are performed to ensure a stable clustering. Finally, the resultant clusters $\{C_1^{\mathrm{EH}}, \ldots, C_K^{\mathrm{EH}}\}$ are provided to the endoscopic expert, who labels each cluster as informative or uninformative. Further processing to define the PSESs is performed only on the informative frames.

The K-means clustering performed on the eigenvectors of the graph Laplacian is also known as spectral clustering [52]. Our method differs from the standard spectral clustering due to the introduction of similarity measures S_{EH} and S_{NCC} instead of using the standard Euclidean Distance, which allows us to adapt the EVMs according to different clustering tasks.

V. Defining the Patient Specific Endoscopic Segments

In the second step of our framework, a new EVM representation is created using only the labelled informative frames of the diagnostic endoscopy. The goal of this clustering step is to group frames showing the same location in the oesophagus into one PSES. Thus, the similarity measure used to create EVMs must highlight the correlation between the visual appearances of two frames. To this end, we use the normalized cross correlation (NCC) similarity measure to define the neighborhood in the manifold learning framework.

A. NCC Similarity Measure

In the original setting of the Laplacian Eigenmaps method, involved in computing the EVMs, the standard choice of the distance (dissimilarity) measure is the Euclidean distance [28] which is equivalent to the sum of squared distances (SSD). The Euclidean distance is a suitable choice for the general manifold learning framework. However, the fact that our input data is limited to endoscopic frames allows us to define a more specific similarity measure; i.e., NCC. The NCC reveals the correlation between two frames and is invariant to linear changes in intensities as opposite to the standard Euclidean distance. The NCC measure is defined as

$$\mathrm{NCC}(\mathcal{I}_i, \mathcal{I}_j) = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \frac{(\mathcal{I}_i(x,y) - \overline{\mathcal{I}_i})(\mathcal{I}_j(x,y) - \overline{\mathcal{I}_j})}{\sigma_{\mathcal{I}_i} \sigma_{\mathcal{I}_j}} \quad (16)$$

where $\overline{\mathcal{I}_i}$, $\overline{\mathcal{I}_j}$ and $\sigma_{\mathcal{I}_i}$, $\sigma_{\mathcal{I}_j}$ denote the mean and standard deviation of the intensity values of images $\mathcal{I}_i$ and $\mathcal{I}_j$, respectively.

Computing the NCC between two images is equivalent to evaluating the following inner product kernel on the image vectors:

$$\mathrm{NCC}(\mathcal{I}_i, \mathcal{I}_j) = \langle \phi_{\mathrm{NCC}}[\mathcal{I}_i], \phi_{\mathrm{NCC}}[\mathcal{I}_j] \rangle \quad (17)$$

with

$$\phi_{\mathrm{NCC}}[\mathcal{I}_i](x,y) = \frac{\mathcal{I}_i(x,y) - \overline{\mathcal{I}_i}}{\|\mathcal{I}_i(x,y) - \overline{\mathcal{I}_i}\|}. \quad (18)$$

In this dot product formulation one can also see that the NCC is equal to the cosine similarity measured on normalized vectors. The cosine similarity is defined based on the angle between two vectors and is therefore a metric inducing a topology on the dataset.

B. Clustering on the EVMs

We include the NCC similarity measure into the EVM framework by choosing $S(\mathcal{I}_i, \mathcal{I}_j) = S_{\mathrm{NCC}}(\mathcal{I}_i, \mathcal{I}_j) = \mathrm{NCC}(\mathcal{I}_i, \mathcal{I}_j)$ (Section III) to compute the new low dimensional representation of the informative frames. Then the K-means clustering [51] is performed in this representation and each cluster in $\{C_1^{\mathrm{PSES}}, \ldots, C_\beta^{\mathrm{PSES}}\}$ is defined as one PSES.

VI. Classifying New Frames

A. Projection Onto the EVMs

For nonlinear manifold learning techniques, such as [28], the mapping from the high to the low dimensional space is not readily extendible to new data points. For the Laplacian Eigenmaps method [28], a linear approximation of this mapping can be computed using the LPP [50], as derived in Section III-F.

Given the informative frames of the diagnostic endoscopy $\hat{\mathcal{I}} = \{\hat{\mathcal{I}}_1, \ldots, \hat{\mathcal{I}}_p\} \subset \mathcal{I}$ and a new surveillance endoscopic frame, we first estimate the optimal linear mapping $\nu : \mathbb{R}^{(w \times h)} \rightarrow \mathbb{R}^d$ from the high dimensional original data space onto the low dimensional representation as explained in Section III-F. To this end, the transformation matrix $\mathbf{T}$ is computed by solving (10). We then project each data point $\mathcal{I}_i$ in the training dataset $\mathcal{I}$ into the low dimensional space using this linear transformation as

$$\mathbf{v}_i = \nu(\mathcal{I}_i) = \mathbf{T}^{\mathsf{T}} \mathcal{I}_i. \quad (19)$$

For the online classification, the new endoscopic frame $\mathcal{I}_s \notin \mathcal{I}$ is projected onto the same low dimensional manifold as $\mathbf{v}_s = \nu(\mathcal{I}_s) = \mathbf{T}^{\mathsf{T}} \mathcal{I}_s$ and the classification is performed in this low dimensional manifold representation using a NN-classification.

B. Assigning a PSES to a Query Frame

In the final classification, we classify a new frame $\mathcal{I}_s \notin \mathcal{I}$ as informative/uninformative and assign it to a PSES (if informative). To do so, we consider as classes the eliminated uninformative clusters $\overline{\Omega^{\mathrm{EH}}} = \{\overline{C_1}^{\mathrm{EH}}, \ldots, \overline{C_\alpha}^{\mathrm{EH}}\}$ together with the PSESs $\Omega^{\mathrm{PSES}} = \{C_1^{\mathrm{PSES}}, \ldots, C_\beta^{\mathrm{PSES}}\}$. Thus, the final set of classes to assign is formed as $\Omega = \{\overline{\Omega^{\mathrm{EH}}}, \Omega^{\mathrm{PSES}}\}$. To classify a new frame $\mathcal{I}_s$, its NN is found among the frames of the diagnostic endoscopy in the low dimensional representation $\nu(\mathcal{I}_s)$.

The class of the NN($\mathcal{I}_s$) is also assigned to the surveillance endoscopic frame $\mathcal{I}_s$. If one of the eliminated uninformative clusters is assigned to a new frame during this online classification, the frame is also considered to be uninformative. If on the other hand one of the PSESs labels is assigned to a new frame, the frame is considered to belong to that PSESs of the diagnostic endoscopy. Once a PSESs containing an optical biopsy location is assigned to a new surveillance endoscopy frame, the endoscopist can be notified online during the examination.

Classification of a new frame into a PSES containing an optical biopsy frame from the diagnostic endoscopy, leads to a frame-level recognition of the previous optical biopsy sites. Once a new endoscopic frame is recognized as containing an optical biopsy location of the diagnostic endoscopy, point-based localization of the optical probe can be achieved by several methods [8]–[11].

VII. Experiments and Results

Experiments are conducted on three upper GI narrow-band endoscopic videos consisting of 1834, 3445, and 1546 frames, respectively. The datasets were acquired by an endoscopic expert at three different upper GI-endoscopic procedures using a state-of-the-art Olympus flexible endoscope. Each of the three steps of the proposed framework is evaluated individually on these datasets. Sections VII-A–VII-C present a quantitative evaluation of the experiments for clustering uninformative frames (Section VII-A), defining PSES (Section VII-B) and classifying a new frame (Section VII-C). Further quantitative evaluation of the overall framework is presented is Section VII-D.

A. Clustering and Labelling Uninformative Frames

In these experiments, we demonstrate the effectiveness of the uninformative frame clustering by evaluating the agreement between the obtained clusters and a ground truth labelling. To this end, we compute the sensitivity[2] and positive prediction value (PPV)[3] of the uninformative frame labelling with respect to the parameters of the created EVMs.

Uninformative frames of each dataset are clustered independently on the $\mathrm{EVM}_{\mathrm{EH}}$ created using the energy histogram S_{EH} similarity matrix as explained in Section IV. In the proposed workflow, the resultant clustering is presented to the endoscopic expert during postprocessing of the diagnostic dataset and clusters chosen by the expert are labelled as uninformative. The ground truth labelling of the uninformative frames is performed manually by the expert for all endoscopic videos. In order to avoid any subjective effect of such a supervision in the quantitative evaluation of the clustering, we define the ground truth label of a cluster as uninformative if it contains more than 50% uninformative frames. It is important to note that is automatic labelling of the clusters with 50% or more uninformative frames is performed for evaluation purposes only. In the clinical workflow, the endoscopic expert selects the uninformative clusters (not the frames).

[2]Sensitivity is also known as recall.

[3]Positive prediction value (PPV) also known as precision.

The proposed S_{EH} similarity yields well structured manifolds, where informative and uninformative frames are well separated as shown in Fig. 6(a), (c), and (e). An example of the results with six clusters is illustrated in Fig. 6(b) and (f).

For quantitative analysis, the sensitivity, PPV and F-measure quality measures of each clustering are evaluated over a varying number of clusters from 2 to 50. Given a clustering $\mathbf{C}_\kappa^{\mathrm{EH}} = \{C_1^{\mathrm{EH}}, \ldots, C_\kappa^{\mathrm{EH}}\}$ with $\kappa \in [2, \ldots, 50]$, true positives; i.e., number of correctly labelled uninformative frames, false positives; i.e., number of incorrectly labelled informative frames and false negatives; i.e., number of uninformative frames labelled as informative, are estimated. Sensitivity, PPV and F-measure are computed as follows:

$$\mathrm{Sensitivity}(\mathbf{C}_\kappa^{\mathrm{EH}}) = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \quad (20)$$

$$\mathrm{PPV}(\mathbf{C}_\kappa^{\mathrm{EH}}) = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \quad (21)$$

$$\mathrm{F}(\mathbf{C}_\kappa^{\mathrm{EH}}) = \frac{2 \cdot \mathrm{PPV}(\mathbf{C}_\kappa^{\mathrm{EH}}) \cdot \mathrm{Sensitivity}(\mathbf{C}_\kappa^{\mathrm{EH}})}{\mathrm{PPV}(\mathbf{C}_\kappa^{\mathrm{EH}}) + \mathrm{Sensitivity}(\mathbf{C}_\kappa^{\mathrm{EH}})}. \quad (22)$$

1) Parameter Selection: In order to evaluate the effect of the EVM parameters on the accuracy of the uninformative frame clustering, all three measures are evaluated over a range of manifold nearest neighbors $k \in [4, \ldots, 20]$, manifold dimensions $d \in [2, \ldots, 20]$ and number of histogram bins used in the energy histograms $b \in [10, \ldots, 160]$ while changing the number of clusters from 2 to 50 $\kappa \in [2, \ldots, 50]$. Fig. 7 illustrates F-measures of uninformative frame labelling on EVMs created with all evaluated parameter values. The F-measure combines the PPV and sensitivity values of the labelling into one quality measure and takes values in $[0, \ldots, 1]$. For the best sensitivity and PPV values of a labelling, F-measure equals to 1. The consistency of the high F-measure values is reflected in the flat F-measure surface-plots over a range of the parameter values for d and k demonstrates that there is not much influence of the particular choice of the parameter values of the proposed method. In regard to the number of histogram bins b, the S_{EH} similarity measure becomes more sensitive as b increases. As shown in Fig. 7(g) and (i), 30 histogram bins result in near to optimum F-measure values for all three datasets.

Sensitivity-PPV plots of all clustering results computed on EVMs with dimensionality $d = 3$, $d = 6$, and $d = 9$ are demonstrated in Fig. 8(a)–(c). Similarly, the sensitivity-PPV values estimated from EVMs with $k = 6$, $k = 10$, and $k = 14$ manifold nearest neighbors are shown in Fig. 8(d)–(f). The large overlap of sensitivity-PPV curves estimated from EVMs with different dimensionality and different neighborhoods of the manifolds illustrates the robustness of the performed clustering on EVMs to the choice of these parameters and its stability for a large range of number of clusters κ. As each cluster will be interactively chosen to be informative or uninformative by the endoscopic expert, smaller number of clusters are desired in the clinical workflow. Fig. 8(g)–(i) demonstrates the effect of the number of histogram bins b used in the energy histograms, where the chosen number of 30 bins leads to slightly improved sensitivity-PPV

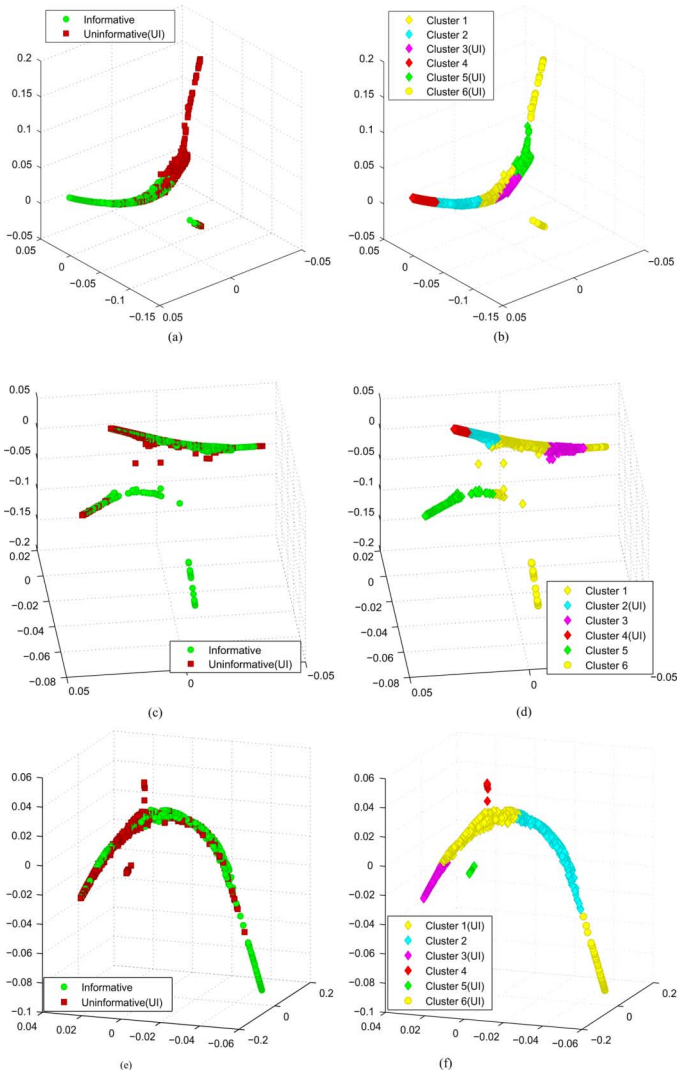

Fig. 6. (a), (c), and (e) First three dimensions of the 6-D EVMs of first, second, and third endoscopic video computed with the proposed S_{EH} similarity measure, respectively. The red squares illustrate the uninformative frames in the ground truth labelling. (b), (d), and (f) Clustering results on the EVMs for the first, second, and third endoscopic video, respectively. The use of S_{EH} in manifold learning leads to structured EVMs where the uninformative frames are clustered together.

values. Finally, we also evaluate the threshold that is used for determining the ground truth informative/uninformative clusters. Fig. 8(g)–(i) shows that if the threshold is smaller than 50% the evaluation becomes more sensitive to the presence of the informative frames within a cluster but its PPV decreases as expected, whereas the contrary holds true for larger threshold values. In our quantitative evaluation, we chose the threshold as 50% that leads to a binary decision without any bias towards informative or uninformative clusters.

2) Comparison of Similarity Measures: In these experiments we demonstrate the accuracy of the proposed energy histogram measure S_{EH} in comparison to a commonly used histogram distance measure; i.e., the Bhattacharyya distance. To this end, the F-measure of both distance measures is computed quantitatively for a range of manifold nearest neighbors $k \in [4, \ldots, 20]$ and dimensionality $d \in [2, \ldots, 20]$ while varying the number of clusters from 2 to 50 for the three datasets. For the same parameters, we also evaluate the Parzen windowing of the energy histograms in order to compare the naive discretization of the energy histogram (EH) to a more advanced estimation of the energy distribution (ED). By approximating the energy distributions with Parzen windowing, we use 30 samples (equal to the number of bins used in energy histograms), 100 samples and 448 samples (equal to the number of frequencies in the rotation-invariant energy histograms $\mathcal{F}_i(\omega_x)$ before discretization into 30 bins). Table I summarizes the performance of all four methods; i.e., cosine and Bhattacharyya distances on EH and ED, on the three datasets.

We have observed that the cosine measure used in S_{EH} in (15) separates the informative and uninformative frames better compared to the Bhattacharyya distance. Furthermore, energy histograms as defined in (14) lead to higher F-measures compared to smoother energy distributions. A possible explanation is that a simple discretization allows for a better discrimination of the energy differences in the high frequencies between informative and uninformative frames in comparison smoother energy distributions. Thus, the proposed measure S_{EH} combining the cosine similarity measure with the energy histograms leads to more accurate clustering of the uninformative frames in comparison to Bhattacharyya distances and energy distributions.

B. Defining Patient Specific Endoscopic Segments

After eliminating the uninformative frames of the endoscopic video, PSESs are defined by performing K-means clustering on the EVMs created from the remaining informative frames as described in Section V. In the following experiments, the separation and compactness of the clusterings of PSESs performed on EVMs are evaluated for different parameter values and a comparison to original image representation and principal component analysis is presented.

1) Parameter Selection: In this study, we consider the manifold dimensionality d and the number of clusters κ to be two parameters of our method and evaluate their effect on the clustering accuracy in relation to each other. For a quantitative evaluation of the values of these two parameters, we compute the Davis–Bouldin index (DB-index) [53] for each clustering while varying the manifold dimensionality from 1 to 40 and the number of clusters from 2 to 40.

Given a clustering $\mathbf{C}_\kappa = \{C_1, \ldots, C_\kappa\}$ with κ clusters, first the within cluster distances (WCDs) are computed as

$$\begin{aligned} \mathrm{WCD}(\mathbf{C}_\kappa) &= [\mathrm{WCD}(C_1), \ldots, \mathrm{WCD}(C_\kappa)]^{\mathsf{T}} \\ \mathrm{WCD}(C_i) &= \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \overline{\mathbf{z_i}}\| \end{aligned} \tag{23}$$

where $\overline{\mathbf{z_i}}$ denotes the center of cluster C_i and n_i denotes the number of elements in cluster C_i. WCD indicates the compactness of the clusters (the smaller the WCDs, the more compact the clusters) and is evaluated in relation to the separability of the clusters which is measured by the between cluster distances (BCDs)

$$\begin{aligned} \mathrm{BCD}(\mathbf{C}_\kappa) &= [\mathrm{BCD}(C_1), \ldots, \mathrm{BCD}(C_\kappa)]^{\mathsf{T}} \\ \mathrm{BCD}(C_i) &= \sum_{j=1, j \neq i}^{\kappa} \|\overline{\mathbf{z_i}} - \overline{\mathbf{z_j}}\|. \end{aligned} \tag{24}$$

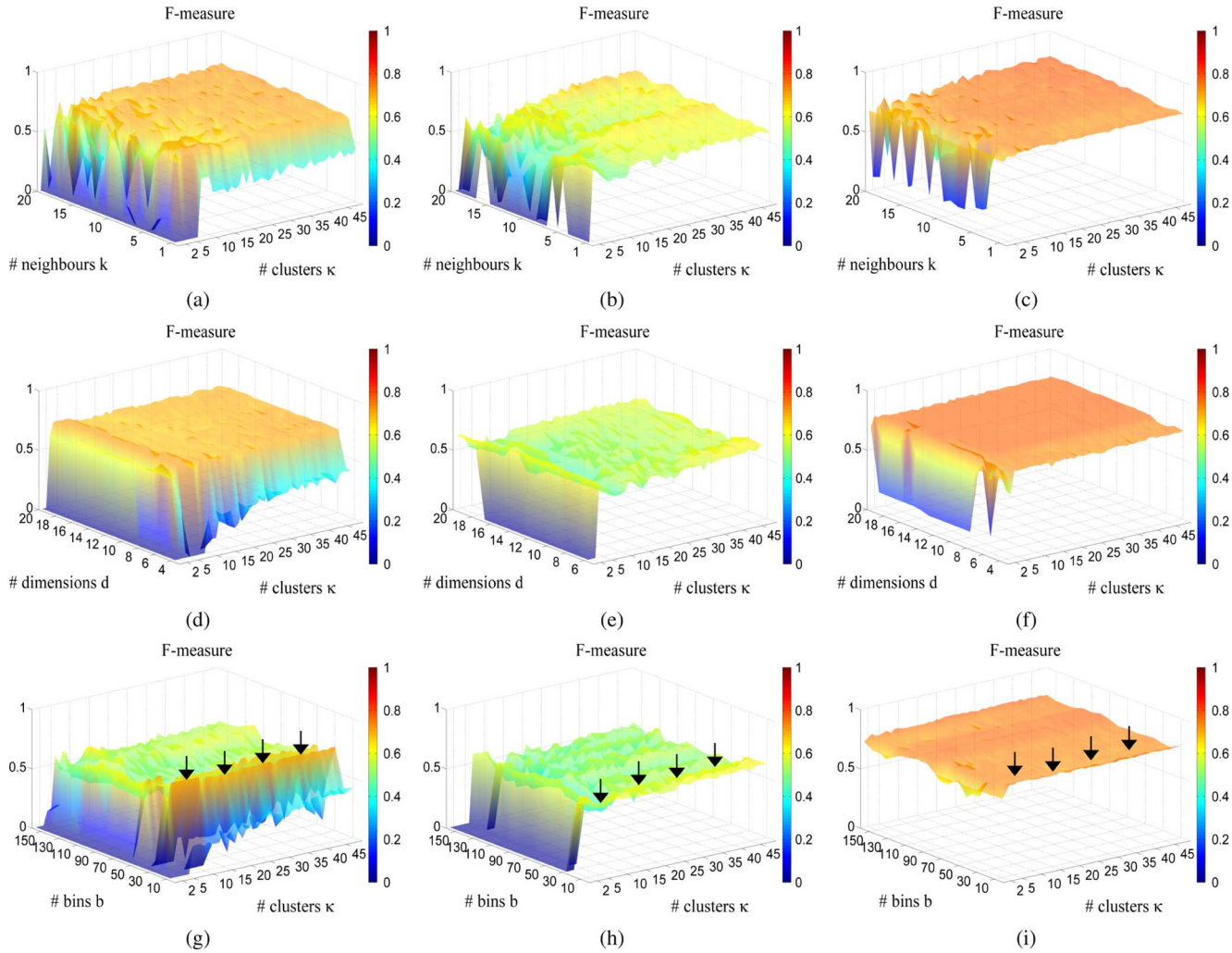

Fig. 7. Evaluation of the EVM parameters for clustering uninformative frames. Sensitivity, PPV and F-measure values are evaluated for different number of manifold nearest neighbors k and dimensions d while changing the number of clusters. (a)–(c) F-measure values as surface-plots for d ranging from 1 to 20 with $k = 6$ and $b = 30$ for the first, second, and third dataset, respectively. (d)–(f) F-measure values for number of manifold neighbors $k \in [4, \ldots, 20]$ and used in creating EVMs with $d = 3$ and $b = 30$ versus different number of clusters κ from 2 to 50 for the three datasets. The flatness of the surface-plots demonstrates the robustness of the sensitivity, PPV and therefore F-measure values to the change of these parameters. (g)–(i) F-measure values for different number of bins $b \in [10, \ldots, 160]$ used in energy histograms and the number of clusters $\kappa \in [2, \ldots, 50]$ on EVMs with $k = 6$ and $d = 3$, where the arrows point the F-measure values for 30 bin energy histograms as used in the rest of this study.

The DB-index is computed as

$$\mathrm{DB}(\mathbf{C}_\kappa) = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \sum_{j=1, j\neq i}^{\kappa} \max \left(\frac{\mathrm{WCD}(C_i) + \mathrm{WCD}(C_j)}{\|\overline{\mathbf{z_i}} - \overline{\mathbf{z_j}}\|} \right). \tag{25}$$

The DB-index is a commonly used evaluation criteria for clustering algorithms and measures the relation of the similarities (or equivalently distances) between clusters and within clusters. This measure is independent of the number of clusters analyzed and its value only depends on the appropriateness of the clustering, which is related to the actual number of clusters in the dataset [53]. Therefore, DB-index allows for the comparison of clusterings with different number of clusters. Smaller DB-indices are desired as they indicate low within cluster distances and high between cluster distances.

The DB-indexes for datasets 1, 2, and 3 are shown in Fig. 9(a), (b), and (c), respectively. In agreement with the study in [52], we observe a significant decrease in the DB-index when the number of clusters is equal or greater the manifold dimensionality, which is clearly reflected in the decrease at the diagonal of DB-index surface-plots in Fig. 9. In the remaining of the experiments, we choose the number of clusters twice the manifold dimensionality: $\kappa = 2 \times d$. This assures a low DB-index and thus more compact and better separated clusters on the EVMs.

2) Comparison of Data Representations: To demonstrate that the EVM representation favours the clustering for defining PSESs, we conduct experiments comparing clusterings in the original image representation, its principal components and EVM representations.

In order to evaluate the quality of the clustering, two criteria are measured. First, the compactness and separability (CS) measure of the clusters is measured based on the within- and between-cluster distances and second DB-index [eq. (25)] is evaluated to measure the separation between clusters. EVM clusterings are compared to K-means clustering performed on the

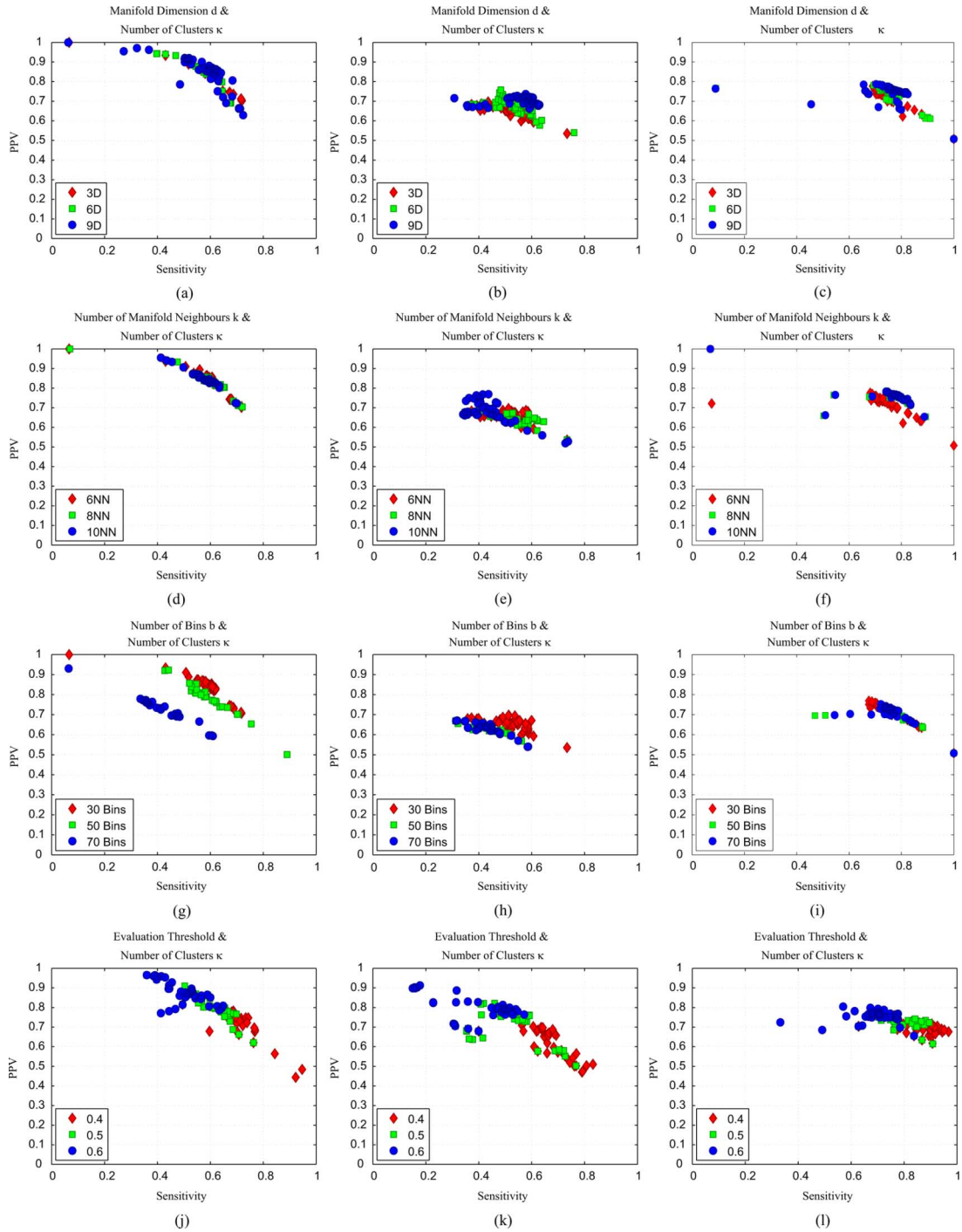

Fig. 8. (a)–(c) Sensitivity versus PPV plots of the clusterings with number of clusters κ ranging between 2 and 50, where the clusterings are performed on $\mathrm{EVM}_{\mathrm{EH}}$ with $d = 3$, $d = 6$ and $d = 9$ ($b = 30$ and $k = 6$) computed using the S_{EH} similarity matrix from the first, second, and third dataset, respectively. (d)–(f) Sensitivity-PPV plots of the clusterings with number of clusters κ ranging between 2 and 50, where the clusterings are performed on EVMs created with $k = 6$, $k = 10$ and $k = 14$ manifold nearest neighbors ($d = 3$ and $b = 30$) from the first, second, and third dataset, respectively. (g)–(i) Sensitivity versus PPV plots for the first, second, and third datasets where the number of bins used in the energy histograms is changed as $b = 30$, $b = 50$, and $b = 70$ while choosing $k = 6$ and $d = 3$. (j)–(l) Sensitivity-PPV plots of the clusterings in relation to the threshold used to label the uninformative clusters in the experiments. The threshold values are varied as 0.4, 0.5, and 0.6 with $k = 6$ and $b = 30$. Note that this threshold is only used to automatically determine the uninformative clusters for evaluation purposes. This step will be replaced by the manual selection of an endoscopic expert in the actual workflow.

original images and on the linear manifold computed using principal component analysis (PCA). The six compared data representations will be referred as follows in the rest of the paper:

- ImSp: Original gray scale images (rescaled to 64×64 pixels resulting in 4096 dimensional data points),
- PCA: Linear manifold of the endoscopic data computed using PCA, where dimensionality is estimated using the method in [54],
- $\mathrm{EVM}_{\mathrm{ED}}$: Nonlinear manifold computed using Laplacian Eigenmaps with the standard Euclidean distance measure,

TABLE I
F-MEASURE VALUES OF COSINE AND BHATTACHARYYA DISTANCE MEASURES EVALUATED ON ENERGY HISTOGRAMS AND ENERGY DISTRIBUTIONS (COMPUTED USING PARZEN WINDOWING) ON THE THREE DATASETS. EVMS FOR BOTH MEASURES; I.E., COSINE AND BHATTACHARYYA DISTANCES ARE CREATED BY RANGING THE PARAMETERS $k \in [6, \ldots, 20]$, $d \in [1, \ldots, 20]$, AND $\kappa \in [2, \ldots, 50]$

		Energy Histograms **30 bins**	***Energy Distribution*** **30 samples**	***Energy Distribution*** **100 samples**	***Energy Distribution*** **448 samples**
		Number of Manifold Neighbours			
Data 1	*Cosine*	0.66 ± 0.05 **(max: 0.72)**	0.36 ± 0.02 (max: 0.51)	0.42 ± 0.01 (max: 0.48)	0.41 ± 0.01 (max: 0.48)
	Bhattacharyya	0.48 ± 0.01 (max: 0.55)	0.22 ± 0.03 (max: 0.42)	0.15 ± 0.03 (max: 0.45)	0.15 ± 0.03 (max: 0.45)
Data 2	*Cosine*	0.52 ± 0.06 **((max: 0.63)**	0.21 ± 0.03 (max: 0.46)	0.15 ± 0.02 (max: 0.36)	0.15 ± 0.02 (max: 0.36)
	Bhattacharyya	0.42 ± 0.01 (max: 0.51)	0.43 ± 0.02 (max: 0.51)	0.38 ± 0.01 (max: 0.44)	0.38 ± 0.01 (max: 0.44)
Data 3	*Cosine*	0.76 ± 0.01 **((max: 0.79)**	0.65 ± 0.02 (max: 0.71)	0.65 ± 0.01 (max: 0.68)	0.65 ± 0.01 (max: 0.69)
	Bhattacharyya	0.63 ± 0.01 (max: 0.69)	0.66 ± 0.03 (max: 0.71)	0.66 ± 0.02 (max: 0.72)	0.66 ± 0.02 (max: 0.72)
		Manifold Dimensionality			
Data 1	*Cosine*	0.64 ± 0.05 **((max: 0.74)**	0.48 ± 0.02 (max: 0.67)	0.53 ± 0.02 (max: 0.70)	0.53 ± 0.02 (max: 0.70)
	Bhattacharyya	0.52 ± 0.05 (max: 0.62)	0.40 ± 0.03 (max: 0.60)	0.39 ± 0.04 (max: 0.60)	0.39 ± 0.04 (max: 0.59)
Data 2	*Cosine*	0.55 ± 0.06 **((max: 0.70)**	0.31 ± 0.04 (max: 0.56)	0.37 ± 0.04 (max: 0.60)	0.36 ± 0.04 (max: 0.59)
	Bhattacharyya	0.41 ± 0.05 (max: 0.53)	0.41 ± 0.05 (max: 0.50)	0.36 ± 0.03 (max: 0.47)	0.34 ± 0.03 (max: 0.50)
Data 3	*Cosine*	0.74 ± 0.01 **((max: 0.80)**	0.68 ± 0.02 (max: 0.74)	0.68 ± 0.02 (max: 0.75)	0.67 ± 0.02 (max: 0.75)
	Bhattacharyya	0.65 ± 0.04 (max: 0.75)	0.65 ± 0.04 (max: 0.72)	0.65 ± 0.04 (max: 0.73)	0.64 ± 0.04 (max: 0.72)

- $\mathrm{EVM}_{\mathrm{NCC}}$: Nonlinear manifold computed using Laplacian Eigenmaps with the NCC measure,
- $\mathrm{vtEVM}_{\mathrm{ED}}$: Visual-temporal manifold computed using Laplacian Eigenmaps with the standard Euclidean distance measure including the temporal constraints, as described in Section III-C,
- $\mathrm{vtEVM}_{\mathrm{NCC}}$: Visual-temporal manifold computed using Laplacian Eigenmaps with the NCC measure including the temporal constraints as described in Section III-C.

The CS-measure is defined as the ratio of the minimum BCD to the average WCD

$$\mathrm{CS}(\boldsymbol{C}_\kappa) = \frac{\min(\mathrm{BCD}(\boldsymbol{C}_\kappa))}{\frac{1}{\kappa}\sum_{c=1}^{\kappa} \mathrm{WCD}(C_c)}. \tag{26}$$

Fig. 10(a), (c), and (e) shows the evaluation of CS-measure for different number of clusters ranging from 5 to 50 for all six representations. For all three endoscopic datasets, clustering on EVMs lead to larger CS values indicating that with this representation clusters become more compact and better separated. A slight decrease in the CS-values is observed when the temporal constraints are included. This can be explained by the imposed temporal continuity of the vtEVMs. As temporally close frames are enforced to be neighbors on the manifold, even if they do not have a high visual similarity, this constraint leads to more continuous manifolds with smaller gaps between the clusters. Thus, clusters on vtEVMs are slightly less separated compared to the EVMs without temporal constraints.

Secondly, we evaluate the DB-index of each clustering [53]. Fig. 10(b), (d), and (f) shows the evaluation of DB-index for different number of clusters ranging from 5 to 50 for $\mathrm{EVM}_{\mathrm{NCC}}$, $\mathrm{EVM}_{\mathrm{ED}}$, $vt\mathrm{EVM}_{\mathrm{NCC}}$, $vt\mathrm{EVM}_{\mathrm{ED}}$, original image and PCA representations. Smaller DB-indexes of clusterings on EVMs demonstrate again that the proposed representation allows for more suitable clustering of the data for all three endoscopic datasets, whereas only slight improvement is observed by using the S_{NCC} instead of the S_{ED}.

Fig. 11(d), (e), and (f) demonstrates sample frames from each cluster on the $\mathrm{EVM}_{\mathrm{NCC}}$ of all three datasets. As shown, each column, which corresponds to one cluster, contains similar endoscopic scenes with varying viewpoint conditions, whereas a significant visual difference between different classes can be observed. Outlier frames in clusters such as in 11th clusters in Fig. 11(e) or fourth cluster in Fig. 11(f) are observed due to some remaining uninformative frames and can be addressed by increasing the number of clusters on $\mathrm{EVM}_{\mathrm{EH}}$. However, due to the manual selection of the uninformative clusters by an endoscopic expert, increasing the number of clusters will also extend the time needed for this supervision and thus the trade-off should be considered.

The EVM representation respects the nonlinear pairwise relations between data points (frames) while mapping each data point into a low dimensional space. Thus, by construction visually different segments become more separated and more compact in this low dimensional EVM representation compared to the high dimensional initial data representation leading to a more efficient clustering as reflected in Fig. 10.

C. Classifying New Frames

In this section, we provide experiments to evaluate the classification accuracy in relation to different numbers of PSESs and in comparison to different data representations. After defining the PSESs by clustering endoscopic frames on EVMs, classification of new frame into one of these defined PSESs is performed by a simple NN classifier as explained in Section VI. The accuracy of the classification step largely depends on the definition of the PSESs and thus on clustering of the diagnostic dataset. In order to evaluate the classification performance in relation to the clusterings, a leave-one-part-out (LOPO) validation is performed on each of the three training videos. To quantitatively evaluate the classification, each frame of an endoscopic video together with its assigned label (its corresponding PSES) is removed from the training dataset and is used as the new surveillance endoscopic frame. In order to prevent a bias caused by the use of one endoscopic dataset per experiment as much as

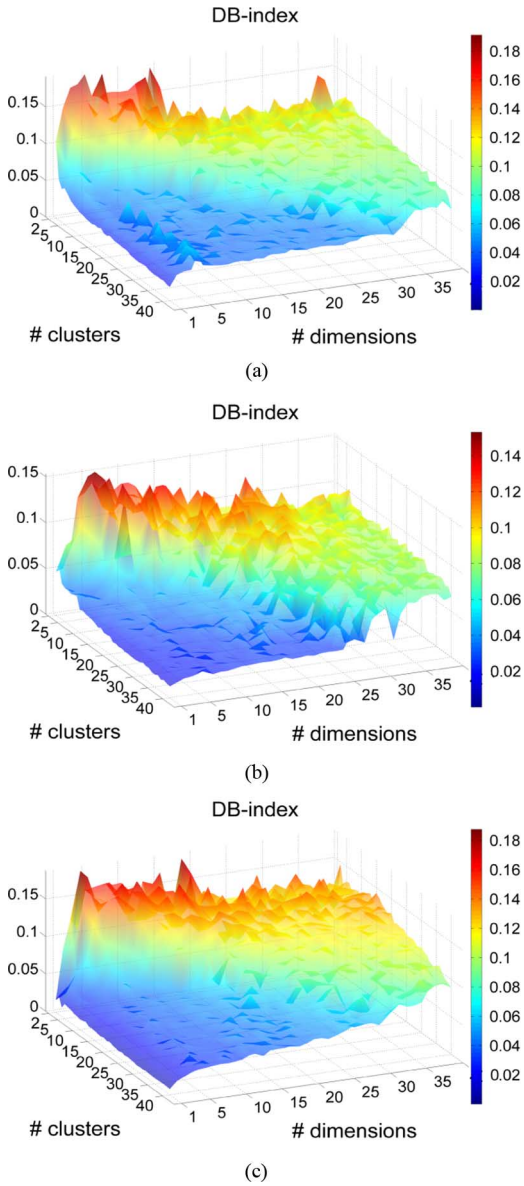

Fig. 9. Evaluation of manifold dimensionality d in relation to the number of clusters κ. DB-indexes are evaluated by varying the number of clusters κ from 2 to 40 and the manifold dimensionality d from 1 to 40. DB-indexes show a significant decrease once the number of clusters is equal or greater the manifold dimensionality.

possible, 40 consecutive frames (20 before and 20 after) are also removed from the training (diagnostic) and test (surveillance) datasets such that the consecutive frames of the test sample are not used in the experiments. This process is repeated sequentially for each frame of the endoscopic datasets. Using the NN classifier as explained in Section VI, one of the PSESs is associated to the new endoscopic frame.

The accuracy of the classification is computed as the ratio of correctly classified frames to all frames of the dataset, whereas the previously known PSES (cluster) of the test frame is used as ground-truth in the comparison. For quantitative evaluation, the LOPO validation is performed for different number of PSESs (clusters used in K-means) ranging from 5 to 50. The classification accuracy with PSESs defined on all six representations are again compared. Fig. 12(a), (c), and (e) shows the classification accuracy for all representations for different number of

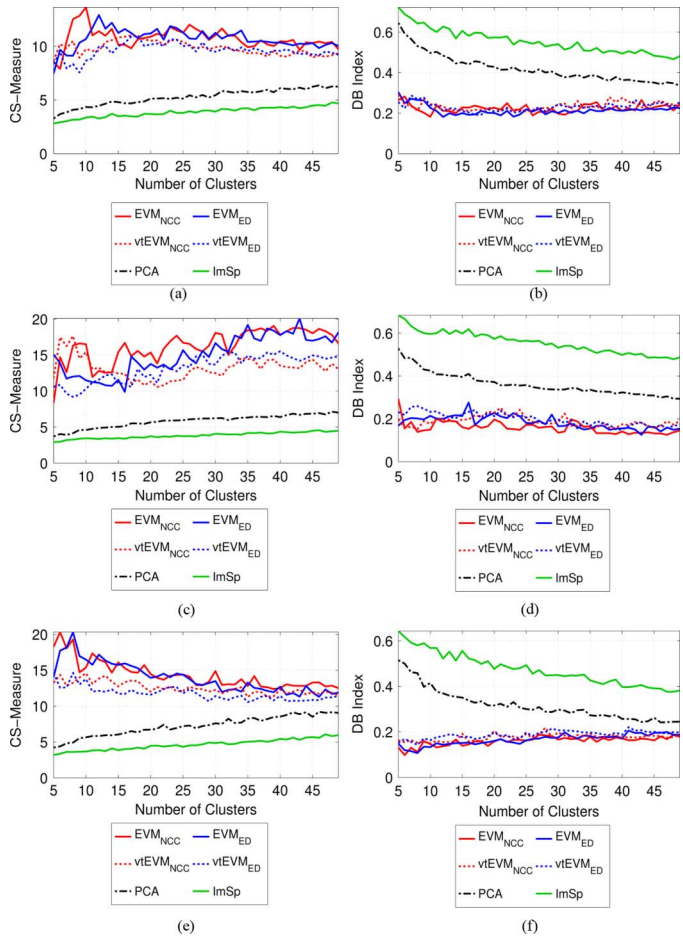

Fig. 10. CS-measure [(a), (c), and (e)] and DB-index [(b), (d), and (f)] evaluated for K-means clustering performed in $\mathrm{EVM_{NCC}}$, $\mathrm{EVM_{ED}}$, $v\mathrm{tEVM_{NCC}}$, $v\mathrm{tEVM_{ED}}$, original image and PCA representations for different number of clusters ranging from 5 to 50 on the first, second, and third endoscopic datasets.

PSESs ranging from 5 to 50. Mean and standard deviation for classification accuracies over all number of clusters is shown in Fig. 12(b), (d), and (f). For all datasets, classification on EVMs leads to higher accuracy compared to the original image and PCA representations. Due to its invariance to linear intensity changes, $\mathrm{EVM_{NCC}}$ results in slightly more accurate classification as compared to the $\mathrm{EVM_{ED}}$.

D. Overall Evaluation

In this final experiment, we perform a quantitative evaluation of the entire proposed framework. We illustrate that using the proposed three steps as combined in our clustering-classification framework yields the highest classification accuracy as compared to performing a direct clustering and classification on the endoscopic videos in the original image space. To this end, the labelled uninformative clusters $\{\overline{C_1}^{\mathrm{EH}}, \ldots, \overline{C_\alpha}^{\mathrm{EH}}\}$ and the PSESs $\{C_1^{\mathrm{PSES}}, \ldots, C_\beta^{\mathrm{PSES}}\}$ are merged as explained in Section VI. Like in the classification experiments in Section VII-C, we perform a LOPO evaluation on the complete endoscopic dataset, this time also including the labelled uninformative frames into the training and the test datasets. The accuracy is computed using the PSESs and as well as the labelled uninformative clusters as classes. The comparison of

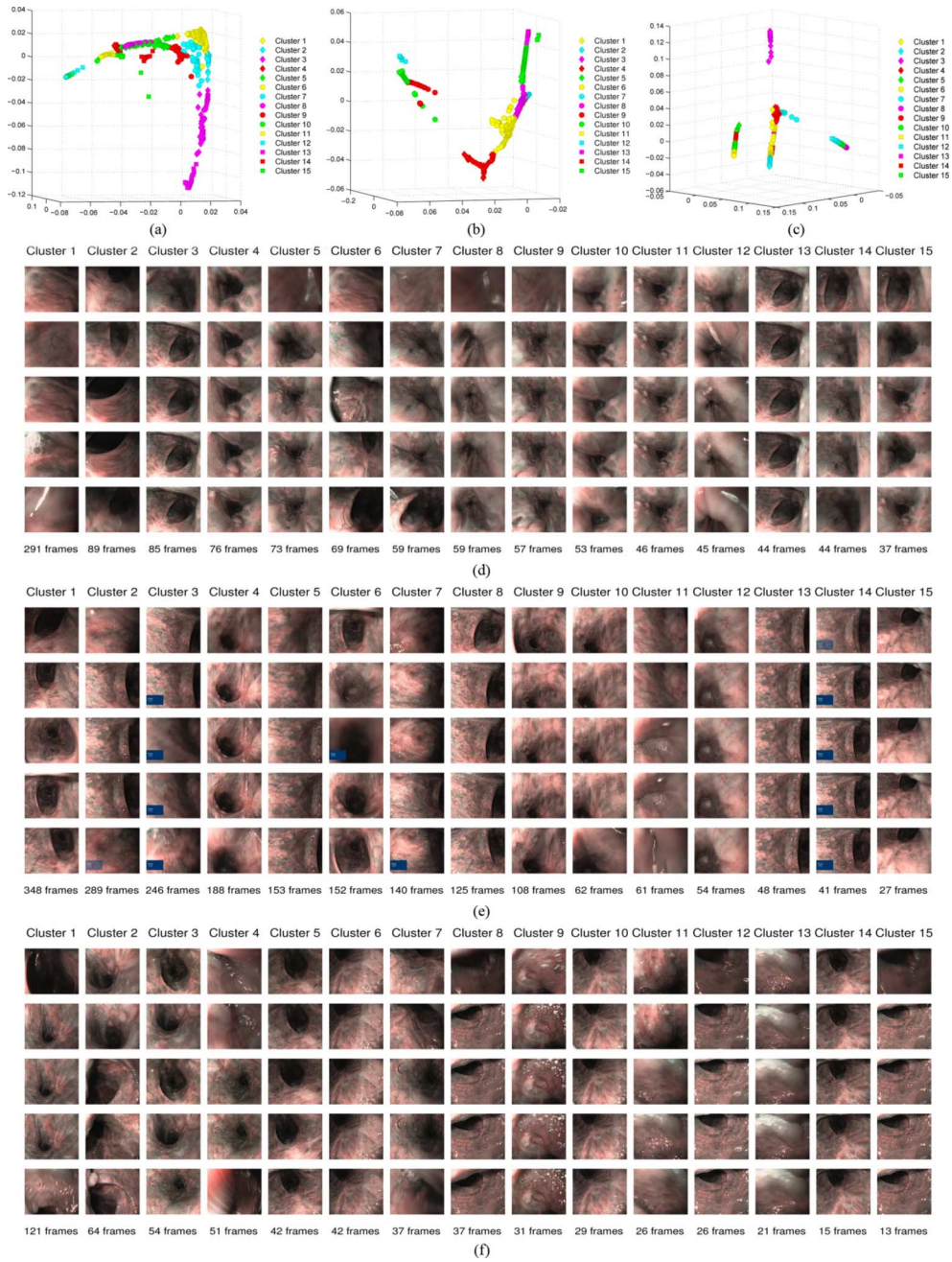

Fig. 11. (a)–(c) First three dimensions of the EVMs computed using the proposed NCC measure after eliminating the uninformative frames from the first, second, and third endoscopic video, respectively. An example clustering performed on $\mathrm{EVM}_{\mathrm{NCC}}$ using K-means algorithm with 15 clusters is demonstrated as color coding. (d)–(f) Results of the clustering on $\mathrm{EVM}_{\mathrm{NCC}}$ for the first, second, and third dataset, respectively. For each dataset all 15 clusters are illustrated, where the first, third, and fifth rows show the first, center and the last frames of each cluster, respectively. Second and fourth row show two example frames of each cluster.

the classification accuracy with number of PSES ranging from 5 to 50 for all three datasets is presented in Fig. 13(a), (c), and (e). The mean and standard deviation of the classification accuracies over the number of PSES is shown in Fig. 13(b), (d), and (f). For all datasets classification on EVMs leads to higher accuracy compared to the original image and PCA representations, whereas different EVMs yield comparable accuracies.

VIII. Discussion and Conclusion

In this work, we introduced a clustering-classification framework to support the retargeting of optical biopsy sites in surveillance endoscopic examinations. Our proposed scheme is based on defining segments of the diagnostic endoscopy in an offline processing step, and on the classification of new frames online during surveillance endoscopy. Offline processing consists of two steps clustering of the endoscopic videos; clustering of first informative/uninformative frames and then of endoscopic segments with different visual appearances. Our method involves supervision in order to allow the expert to select the informative frame clusters. This user interaction is required between the two clustering steps. Online classification is performed using a NN classifier in the introduced EVM representation.

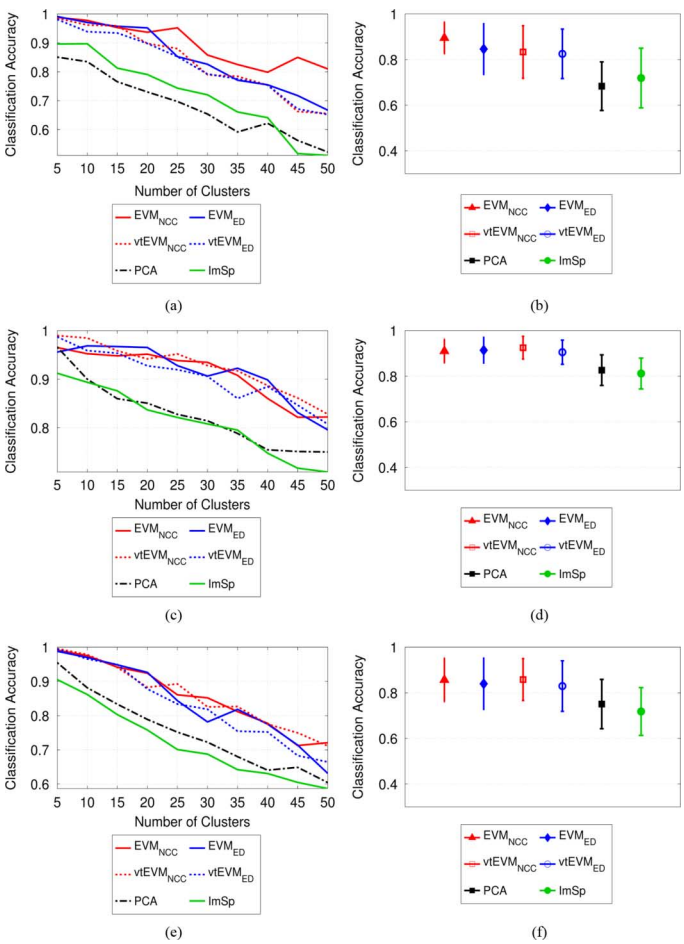

Fig. 12. Classification accuracy estimated by LOPO experiments on the (a) first, (c) second, and (e) third dataset using the PSESs defined on the informative frames only. The experiments are carried out for different number of PSESs ranging from 5 to 50 defined as the clusterings in image space, PCA, $\mathrm{EVM_{ED}}$, $\mathrm{EVM_{NCC}}$, $\mathrm{vtEVM_{ED}}$, and $\mathrm{vtEVM_{NCC}}$. (b), (d), and (f) Mean and standard deviation of classification accuracies over all number of PSESs for the first, second, and third dataset, respectively.

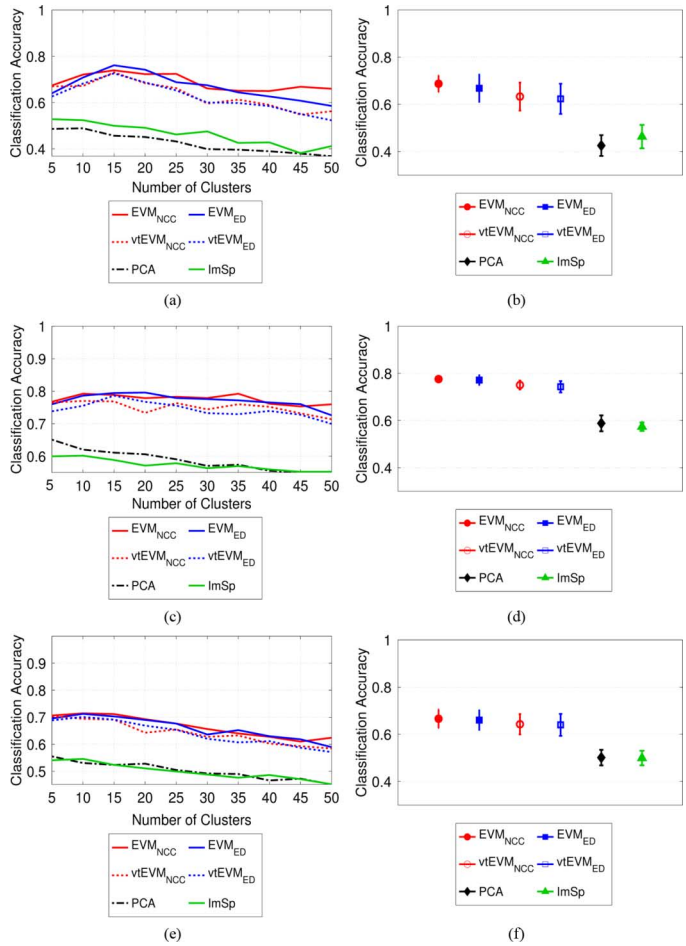

Fig. 13. Classification accuracy estimated by LOPO experiments on the (a) first, (c) second, and (e) third dataset including the uninformative clusters. The experiments are carried out for different number of PSESs ranging from 5 to 50 defined as the clusterings in image space, PCA, $\mathrm{EVM_{ED}}$, $\mathrm{EVM_{NCC}}$, $\mathrm{vtEVM_{ED}}$ and $\mathrm{vtEVM_{NCC}}$. (b), (d), and (f) Mean and standard deviation of classification accuracies over all number of PSESs for the first, second, and third dataset, respectively.

In terms of technical contributions, we investigated the effect of learning the underlying manifold of endoscopic video datasets induced by different notions of similarity on clustering and classification tasks. Taking advantage of the mathematical framework behind manifold learning, different representations of a dataset are created only by redefining the pairwise similarity measure. Each of these EVM representations highlights different aspects of the same dataset allowing for different groupings of the data using standard clustering techniques. Introducing EVMs allows us to address the task of clustering informative frames and endoscopic scenes within the same generic manifold learning framework. Furthermore, we presented a classification approach compatible with this low dimensional representation.

In our experiments, we demonstrated that the proposed EVM representation yields higher accuracy in clustering endoscopic segments as well as in classification of new frames compared to the original image representation and PCA. This improvement is due to the more compact and better separated representation of different clusters of the endoscopic video on the EVMs. In addition, having much lower dimensionality than the original image representation, EVMs provide a means to perform clustering and classification tasks in a more efficient manner. Each step of the proposed framework is evaluated individually on three patient datasets acquired during upper GI-endoscopic examinations. Quantitative analysis in comparison to the original image and PCA representations demonstrates the improved clustering and classification performances achieved on EVMs. We further investigated the effect of including temporal constraints into the EVM framework. All proposed EVM representations; i.e., $\mathrm{EVM_{NCC}}$, $\mathrm{EVM_{ED}}$, $vt\mathrm{EVM_{NCC}}$, and $vt\mathrm{EVM_{ED}}$ show comparable accuracy in our clustering and classification experiments. Further experiments demonstrate the effectiveness of the entire proposed clustering-classification framework quantitatively in LOPO experiments.

Future work will direct towards evaluating the proposed framework on several endoscopic video sequences of the same patient acquired at different time instances. Such evaluation would provide the bridge between the encouraging results of the proposed clustering-classification framework, as presented in this work, and its application in the daily clinical routine of

upper GI-endoscopic examinations to facilitate targeted optical biopsies.

ACKNOWLEDGMENT

The authors would like to thank F. Orihuela-Espina for valuable discussions.

REFERENCES

[1] S. Spechler, "Barrett's esophagus," *GI Motility Online*, 2006.
[2] A. Meining, M. Bajbouj, S. Delius, and C. Prinz, "Confocal laser scanning microscopy for in vivo histopathology of the gastrointestinal tract," *Arab J. Gastroenterol.*, vol. 8, no. 1, pp. 1–4, 2007.
[3] M. Bajbouj, S. Delius, V. Becker, A. Jung, and A. Meining, "Confocal laser scanning endomicroscopy for in vivo histopathology of the gastrointestinal tract and beyond-an update," *Arab J. Gastroenterol.*, vol. 11, no. 4, pp. 181–186, 2010.
[4] B. André, T. Vercauteren, A. Perchant, A. Buchner, M. Wallace, and N. Ayache, "Endomicroscopic image retrieval and classification using invariant visual features," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, 2009, pp. 346–349.
[5] B. André, T. Vercauteren, A. Buchner, M. Wallace, and N. Ayache, "Endomicroscopic video retrieval using mosaicing and visual words," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, 2010, pp. 1419–1422.
[6] B. André, T. Vercauteren, A. Perchant, A. Buchner, M. Wallace, and N. Ayache, "Introducing space and time in local feature-based endomicroscopic image retrieval," in *Proc. Med. Content-Based Retrieval Clinical Decision Support, (MCBR-CDS)*, 2009, pp. 18–30.
[7] B. André, T. Vercauteren, A. Buchner, M. Shahid, M. Wallace, and N. Ayache, "An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MIC-CAI)*, 2010, pp. 480–487.
[8] B. Allain, M. Hu, L. Lovat, R. Cook, S. Ourselin, and D. Hawkes, "Biopsy site re-localisation based on the computation of epipolar lines from two previous endoscopic images," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MIC-CAI)*, 2009, pp. 491–498.
[9] B. Allain, M. Hu, L. Lovat, R. Cook, T. Vercauteren, S. Ourselin, and D. Hawkes, "A system for biopsy site re-targeting with uncertainty in gastroenterology and oropharyngeal examinations," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MIC-CAI)*, 2010, pp. 514–521.
[10] P. Mountney, S. Giannarou, D. Elson, and G. Yang, "Optical biopsy mapping for minimally invasive cancer screening," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MIC-CAI)*, 2009, pp. 483–490.
[11] S. Atasoy, B. Glocker, S. Giannarou, D. Mateus, A. Meining, G. Yang, and N. Navab, "Probabilistic region matching in narrow-band endoscopy for targeted optical biopsy," in *Proc. Int. Conf. Med. Image Computing Comput. Assist. Intervent. (MIC-CAI)*, 2009, pp. 499–506.
[12] M. Bashar, K. Mori, Y. Suenaga, T. Kitasaka, and Y. Mekada, "Detecting informative frames from wireless capsule endoscopic video using color and texture features," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MIC-CAI)*, 2008, pp. 603–610.
[13] M. Bashar, T. Kitasaka, Y. Suenaga, Y. Mekada, and K. Mori, "Automatic detection of informative frames from wireless capsule endoscopy images," *Med. Image Anal.*, vol. 14, no. 3, pp. 449–470, 2010.
[14] D. Iakovidis, S. Tsevas, and A. Polydorou, "Reduction of capsule endoscopy reading times by unsupervised image mining," *Comput. Med. Imag. Graph.*, vol. 34, no. 6, pp. 471–478, 2010.
[15] M. Arnold, A. Ghosh, G. Lacey, S. Patchett, and H. Mulcahy, "Indistinct frame detection in colonoscopy videos," in *Proc. IEEE Int. Mach. Vis. Image Process. Conf. (IMVIP)*, 2009, pp. 47–52.
[16] J. Oh, S. Hwang, J. Lee, W. Tavanapong, J. Wong, and P. deGroen, "Informative frame classification for endoscopy video," *Med. Image Anal.*, vol. 11, no. 2, pp. 110–127, 2007.
[17] T. Stehle, R. Auer, S. Gross, A. Behrens, J. Wulff, T. Aach, R. Winograd, C. Trautwein, and J. Tischendorf, "Classification of colon polyps in NBI endoscopy using vascularization features," *SPIE Med. Imag. Computer-Aided Diagnosis*, 2009.
[18] S. Gross, T. Stehle, A. Behrens, R. Auer, T. Aach, R. Winograd, C. Trautwein, and J. Tischendorf, "A comparison of blood vessel features and local binary patterns for colorectal polyp classification," *SPIE Med. Imag. Computer-Aided Diagnosis*, vol. 7260, 2009.
[19] T. Tamaki, J. Yoshimuta, T. Takeda, B. Raytchev, K. Kaneda, S. Yoshida, Y. Takemura, and S. Tanaka, "A system for colorectal tumor classification in magnifying endoscopic NBI images," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2010, vol. 2, pp. 987–998.
[20] S. Karkanis, D. Iakovidis, D. Maroulis, G. Magoulas, and N. Theofanous, "Tumor recognition in endoscopic video images using artificial neural network architectures," in *Proc. IEEE Euromicro Conf.*, 2002, vol. 2, pp. 423–429.
[21] D. Iakovidis, D. Maroulis, and S. Karkanis, "An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy," *Comput. Biol. Med.*, vol. 36, no. 10, pp. 1084–1103, 2006.
[22] D. Maroulis, D. Iakovidis, S. Karkanis, and D. Karras, "CoLD: A Versatile detection system for colorectal lesions in endoscopy video-frames," *Comput. Methods Programs Biomed.*, vol. 70, no. 2, pp. 151–166, 2003.
[23] S. Karkanis, D. Iakovidis, D. Karras, and D. Maroulis, "Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures," in *Proc. IEEE Int. Conf. Image Process.*, 2002, vol. 2, pp. 833–836.
[24] D. Iakovidis, S. Tsevas, D. Maroulis, and A. Polydorou, "Unsupervised summarisation of capsule endoscopy video," in *Proc. 4th Int. IEEE Conf. Intell. Syst.*, 2008, vol. 1, pp. 3–15.
[25] S. Atasoy, D. Mateus, J. Lallemand, A. Meining, G.-Z. Yang, and N. Navab, "Endoscopic video manifolds," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MIC-CAI)*, Sep. 2010, pp. 437–445.
[26] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, p. 2319, 2000.
[27] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, p. 2323, 2000.
[28] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Computat.*, vol. 15, no. 6, pp. 1373–1396, 2003.
[29] R. Coifman and S. Lafon, "Diffusion maps," *Appl. Computat. Harmonic Anal.*, vol. 21, no. 1, pp. 5–30, 2006.
[30] C. Wachinger, M. Yigitsoy, and N. Navab, "Manifold learning for image-based breathing gating with application to 4D ultrasound," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MIC-CAI)*, 2010, pp. 26–33.
[31] P. Tiwari, M. Rosen, and A. Madabhushi, "Consensus-locally linear embedding (C-LLE): Application to prostate cancer detection on magnetic resonance spectroscopy," in *Proc. Int. Conf. Med. Image Computing Comput. Assist. Intervent. (MIC-CAI)*, 2008, pp. 330–338.
[32] P. Tiwari, M. Rosen, G. Reed, J. Kurhanewicz, and A. Madabhushi, "Spectral embedding based probabilistic boosting tree (ScEPTre): Classifying high dimensional heterogeneous biomedical data," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MICCAI)*, 2009, pp. 844–851.
[33] K. Suzuki, J. Zhang, and J. Xu, "Massive-training artificial neural network coupled with Laplacian-eigenfunction-based dimensionality reduction for computer-aided detection of polyps in CT colonography," *IEEE Trans. Med. Imag.*, vol. 29, no. 11, pp. 1907–1917, Nov. 2010.
[34] R. Sparks and A. Madabhushi, "Novel morphometric based classification via diffeomorphic based shape representation using manifold learning," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MICCAI)*, 2010, pp. 658–665.
[35] Q. Zhang, R. Souvenir, and R. Pless, "On manifold structure of cardiac MRI data: Application to segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2006, vol. 1, pp. 1092–1098.
[36] P. Etyngier, F. Ségonne, and R. Keriven, "Active-contour-based image segmentation using machine learning techniques," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MICCAI)*, 2007, pp. 891–899.
[37] S. Kadoury and N. Paragios, "Nonlinear embedding towards articulated spine shape inference using higher-order MRFs," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MICCAI)*, 2010, pp. 579–586.
[38] M. Georg, R. Souvenir, A. Hope, and R. Pless, "Simultaneous data volume reconstruction and pose estimation from slice samples," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2008, pp. 1–6.
[39] J. Hamm, D. Ye, R. Verma, and C. Davatzikos, "GRAM: A framework for geodesic registration on anatomical manifolds," *Med. Image Anal.*, vol. 14, no. 5, pp. 633–642, 2010.
[40] R. Souvenir and R. Pless, "Image distance functions for manifold learning," *Image Vis. Comput.*, vol. 25, no. 3, pp. 365–373, 2007.

[41] K. Lekadir, D. Elson, J. Requejo-Isidro, C. Dunsby, J. McGinty, N. Galletly, G. Stamp, P. French, and G. Yang, "Tissue characterization using dimensionality reduction and fluorescence imaging," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent. (MICCAI)*, 2006, vol. 4191, p. 586.

[42] L. Schwarz, D. Mateus, and N. Navab, "Multiple-activity human body tracking in unconstrained environments," *Articulated Motion Deformable Objects*, pp. 192–202, 2010.

[43] L. Van der Maaten, E. Postma, and H. V. Herik, Dimensionality reduction: A comparative review Tilburg Univ., Tilburg, The Netherlands, 2009.

[44] R. Pless and R. Souvenir, "A survey of manifold learning for images," *IPSJ Trans. Comput. Vis. Appl.*, vol. 1, pp. 83–94, 2009.

[45] T. Lin and H. Zha, "Riemannian manifold learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 796–809, May 2008.

[46] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2006, vol. 2, pp. 2169–2178.

[47] D. Gorisse, M. Cord, F. Precioso, and S. Philipp-Foliguet, "Fast approximate kernel-based similarity search for image retrieval task," in *Proc. 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[48] B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 12, pp. 2143–2157, Dec. 2009.

[49] P. Indyk and N. Thaper, "Fast image retrieval via embeddings," in *Proc. 3rd Int. Workshop Stat. Computat. Theories Vis.*, 2003.

[50] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[51] J. Hartigan and M. Wong, "A k-means clustering algorithm," *JR Stat. Soc., Ser. C*, vol. 28, pp. 100–108, 1979.

[52] U. VonLuxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[53] D. Davies and D. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 1, pp. 224–227, Apr. 1979.

[54] K. Fukunaga and D. Olsen, "An algorithm for finding intrinsic dimensionality of data," *IEEE Trans. Computers*, vol. 100, no. 2, pp. 176–183, Feb. 1971.