Endoscopic Video Manifolds

Selen Atasoy[1,2], Diana Mateus[1], Joe Lallemand[1], Alexander Meining[3](MD), Guang-Zhong Yang[2], and Nassir Navab[1]

[1] Chair for Computer Aided Medical Procedures (CAMP), TU Munich, Germany
{atasoy,mateus,lalleman,navab}@cs.tum.edu
[2] Visual Information Processing Group, 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. Postprocedural analysis of gastrointestinal (GI) endoscopic videos is a difficult task because the videos often suffer from a large number of poor-quality frames due to the motion or out-of-focus blur, specular highlights and artefacts caused by turbid fluid inside the GI tract. Clinically, each frame of the video is examined individually by the endoscopic expert due to the lack of a suitable visualisation technique. In this work, we introduce a low dimensional representation of endoscopic videos based on a manifold learning approach. The introduced endoscopic video manifolds (EVMs) enable the clustering of poor-quality frames and grouping of different segments of the GI endoscopic video in an unsupervised manner to facilitate subsequent visual assessment. In this paper, we present two novel inter-frame similarity measures for manifold learning to create structured manifolds from complex endoscopic videos. Our experiments demonstrate that the proposed method yields high precision and recall values for uninformative frame detection (90.91% and 82.90%) and results in well-structured manifolds for scene clustering.

Keywords: Endoscopy, manifold learning, video segmentation, clustering.

1 Introduction

GI endoscopy is a widely used clinical technique for visualising the digestive tract. Current diagnosis and surveillance of GI diseases, ranging from Barrett's Oesophagus to oesophageal or colorectal cancer, are performed by visual assessment in GI endoscopy followed by necessary biopsies. Clinically, endoscopic videos also serve the postprocedural analysis performed by the expert and subsequent image processing for quantitative analysis. Currently, postprocedural analysis is typically performed by the endoscopic expert via visual assessment of each frame in the sequence. Such an analysis is complicated and time consuming mainly due to two reasons. First, in a typical endoscopic video sequence, there are usually a large number of poor-quality frames due to the blur caused by fast motion or out-of-focus imaging of the endoscope, specular highlights and artefacts caused by the turbid fluid inside the GI tract (Fig.1). Second, each frame

(a) (b) (c) (d) (e)

Fig. 1. a) illustrates an ideal frame acquired by a state-of-the art GI endoscope. (b-e) show several challenges encountered in endoscopic videos; frames with motion blur b), specular highlights c), bubbles caused by the liquid inside the organ d) and blur caused by out-of-focus e).

in the sequence is inspected individually by the expert, as there exists no easily manageable visualisation technique for GI endoscopic videos.

In endoscopic video analysis, the focus is mainly directed towards detecting abnormalities [1–3] and uninformative frames [4, 5]. These methods focus on defining specific features such as colour or texture and then detecting the frames containing them in order to present the expert only the detected informative frames instead of the whole content of the endoscopic video. Recently, representative frame extraction for content summary has also been investigated to aid the postprocedural analysis of wireless capsule endoscopy [6]. The aim of our work is to cluster the GI endoscopic videos in an *unsupervised* manner in order to allow the *expert* to easily eliminate or visualise only the parts of interest during postprocedural analysis. To this end, we introduce endoscopic video manifolds (EVMs); a low dimensional representation of endoscopic videos based on manifold learning that allows for clustering of different scenes as well as of poor quality frames.

Successful manifold learning algorithms have been proven to be beneficial for a range of image processing tasks, e.g. [7, 8]. The main novelty of these methods in comparison to feature or intensity based image representation techniques lies in analysing a set of images based on their similarities. In [8], Pless proposed a video representation using a low dimensional image space and a trajectory for analysing natural video sequences. In this work, we will explore the use of manifold learning techniques to perform clustering on GI endoscopic videos.

The contribution of this work is twofold: firstly, from the medical point of view, we propose EVMs as a generic approach to cluster poor-quality frames as well as different segments of the GI endoscopic video in an unsupervised manner. This allows the experts to easily analyse the segment of interest. Secondly, in terms of theoretical contribution, we propose two inter-frame similarity measures for manifold learning, namely rotation invariant energy histograms and divergence of the optical flow field, which create structured manifolds from the complex endoscopic scenes. The first measure enhances the spectral differences between an ideal and a poor-quality frame while the second measure leads to closer localisation of similar frames on the manifold by considering temporal constraints among them. The design of these similarity measures is necessary as we are confronted with the difficult imaging conditions of endoscopy.

2 Methods

We address two tasks: clustering of *poor-quality frames* and *endoscopic scenes*. For each task our method creates a manifold representation using an appropriate inter-frame similarity measure and performs a clustering on the created EVM.

2.1 Overview of the Framework

An endoscopic video $\mathcal{I}$ can be represented by the set of its n individual frames $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_n\}$. Each frame is a data point in the high dimensional input space $\mathcal{I}_1, \mathcal{I}_2, \cdots \mathcal{I}_n \in \mathbb{R}^{w \times h}$, where w and h are the width and height of the frames, respectively. Thus, the number of degrees of freedom (DoF) is equal to $w \times h$. However, due to the continuity of the video sequence, and therefore the large similarity between consecutive frames, the actual DoF is much smaller than this discrete representation enables. So, the high dimensional data points actually lie on a lower dimensional manifold $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_n \in \mathcal{M}$, where $\mathcal{M}$ is a manifold embedded in $\mathbb{R}^{w \times h}$. We compute the low dimensional EVM as follows:

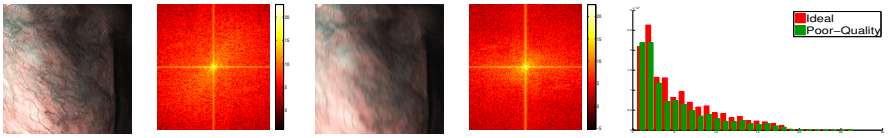
1. Defining the similarities: For each pair $(\mathcal{I}_i, \mathcal{I}_j)$, of the given n data points $i, j \in \{1, \cdots, n\}$, first a similarity measure is defined $W : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$. W determines which images are considered to be similar and therefore kept as neighbours on the manifold. Thus, the similarity measure determines the structure of the manifold and should be designed carefully for each particular application. In the sections 2.2 and 2.3 we present the similarity measures designed for the addressed clustering tasks.

2. Computing the adjacency graph: Given the similarity matrix W, where the values $W(i, j)$ state the similarity between the frames $\mathcal{I}_i$ and $\mathcal{I}_j$, first, k-nearest neighbours of each data point are computed. Then, the adjacency graph is created as:

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

where $\mathcal{N}_j^k$ states the k-nearest neighbours of the j-th data point. Then, a connected component analysis is performed on the adjacency graph and the low dimensional manifold is computed for each component separately.

3. Learning the manifold: In this work, we use the *local* manifold learning based on Laplacian Eigenmaps (LE) [9]. The choice of this local method is driven by the observation that for the GI endoscopic videos distant data points on the manifold (corresponding to non-similar images) do not yield meaningful similarity measures. Therefore, local methods which do not take these similarities into consideration are better suited for our application compared to the global methods as used in [7, 8]. To compute the LE, the eigenvalues and eigenvectors $\{f_1, \cdots, f_m\}$ of the Laplacian matrix $L = D - A$ are determined, where D represents the degree matrix $D(i, i) = \sum_j A(i, j)$. The m-dimensional ($m \ll w \times h$) representation of a frame $\mathcal{I}_i$ on the EVM is then given by $[f_1(i), \cdots, f_m(i)]^\top$.



(a) (b) (c) (d) (e)

Fig. 2. Rotation invariant energy histograms. a) and b) show an ideal frame and its power spectrum, respectively. c) and d) show a blurred frame and its power spectrum, respectively. e) shows energy histograms of the ideal and blurred frames.

4. Clustering on the manifold: Finally, the clustering of the uninformative frames and video segments is performed on the corresponding EVM using the K-means algorithm [10]. Thus, the endoscopic video $\mathcal{I}$ is represented as a set of l clusters $\mathcal{I} = \{C_1, \cdots, C_l\}$. The results of the clustering depends on the structure of the manifold and thus on the chosen similarity measure. As next we present, the construction of the EVMs for the two addressed tasks.

2.2 EVM for Clustering Uninformative Frames

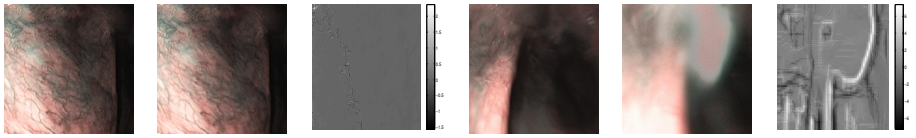
In order to create an EVM, where the poor-quality frames are closely localized, we propose to use a new inter-frame similarity measure based on the power spectrum of the images. In the frequency domain, the energy of an ideal frame is more distributed over low and high frequencies compared to a poor-quality frame whose energy is mainly accumulated only in low frequencies (Fig.2). Therefore, the EVM is created using the inter-frame similarity measure based on rotation invariant energy histograms. To this end, first the power spectrum of a frame $\mathcal{I}_i$ is represented in log-polar coordinates $\mathcal{F}_i(f, \theta)$, where f and θ state the frequencies and the orientations, respectively. Then the rotation invariant power spectrum is computed as: $\mathcal{F}_i(f) = \sum_\theta \mathcal{F}_i(f, \theta)$ and an histogram with B bins $\text{hist}(\mathcal{F}_i(f), B)$ is created. Finally, the EVM is created by using the following similarity measure for all pairs of frames $(\mathcal{I}_i, \mathcal{I}_j)$:

$$W_{\text{EH}}(\mathcal{I}_i, \mathcal{I}_j) = \pi - acos\left(\frac{\left\langle \text{hist}^b(\mathcal{F}_i(f), B), \text{hist}^b(\mathcal{F}_j(f), B)\right\rangle}{||\text{hist}^b(\mathcal{F}_i(f), B)|| \cdot ||\text{hist}^b(\mathcal{F}_j(f), B)||}\right), \quad (2)$$

where hist^b states the b-th bin of the histogram, $\langle \cdot, \cdot \rangle$ is the dot product and $||hist||$ denotes the norm of the B-valued histogram vector. In this paper, we use $b = 30$ for our experiments. However, it is noted that there has not been a significant difference in the manifold structure for different values of n.

2.3 EVMs for Clustering Endoscopic Scenes

For clustering endoscopic scenes, we create two different EVMs; the first one based on the endoscope motion considering the temporal constraints (Sec.2.3a) and the second one considering the appearance similarities of all frames (Sec.2.3b).



(a) (b) (c) (d) (e) (f)

Fig. 3. Divergence of optical flow field. a) and b) show two consecutive frames with ideal conditions and c) illustrates the small and constant divergence field of the optical flow between a) and b). d) and e) show two consecutive frames with different conditions (one ideal and one non-informative frame). e) illustrates the varying divergence field with of the optical flow field between them.

a) Optical Flow Based EVMs: Changes in a GI-endoscopic video are caused mainly by the motion of the endosopce. Therefore a measure of the camera motion indicates directly a change in the observed scene. We propose using the optical flow divergence which measures the smoothness of camera motion field. This measure will lead to a high similarity between two images only if the scene and the imaging conditions (such as blur, specular highlights) are similar (Fig.3) If the optical flow field $\Phi_i^j(x, y)$ from i-th frame ($\mathcal{I}_i$) to j-th frame ($\mathcal{I}_j$) is a smooth motion field, then the divergence at each location will be close to 0. Thus, the similarity between $\mathcal{I}_i$ and $\mathcal{I}_j$ is computed as:

$$W_{\text{DOFF}}(\mathcal{I}_i, \mathcal{I}_j) = 1 - \frac{\psi_i^j}{\max(\psi_i^j)}, \qquad \psi_i^j = \sum_{x=1}^{w} \sum_{y=1}^{h} |\nabla \Phi_i^j(x, y)|, \qquad \Phi_i^j : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}^2 \tag{3}$$

where ∇ is the divergence operator $\nabla = \partial/\partial x + \partial/\partial y$. In order to consider temporal constraints, the k-nearest neighbours of a frame $\mathcal{I}_i$ are searched only within the frames $\{\mathcal{I}_{i-s}, \cdots, \mathcal{I}_{i+s}\}$, where s is the size of the search window (25 frames in this study). For the computation of the optical flow method, we use without loss of generality the optical flow method of Black and Anandan [11].

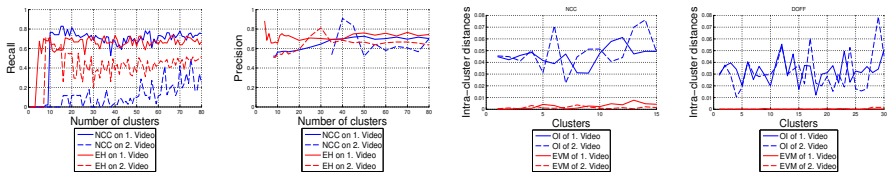
b) Intensity Based EVMs Finally, we also create EVMs using the Normalised cross correlation (NCC) as similarity measure: $W_{\text{NCC}}(\mathcal{I}_i, \mathcal{I}_j) = NCC(\mathcal{I}_i, \mathcal{I}_j)$.

3 Experiments and Results

The experiments are conducted on two upper GI narrow-band endoscopic videos consisting of 1834 and 1695 frames. The datasets are acquired by an endoscopic expert at two different GI-endoscopic procedures. The ground truth labelling of poor-quality frames is performed manually by the expert for both videos.

3.1 Clustering Uninformative Frames

For this task, two EVMs are created using W_{EH} and W_{NCC} similarity measures. For quantitative analysis, *recall* and *precision* values of each clustering are evaluated over the number of clusters from 1 to 80. After clustering on the EVMs,



(a) (b) (c) (d)

Fig. 4. a)Recall and b)precision values for clustering uninformative frames on EVMs created with W_{EH} and W_{NCC}. Note that for the 2. dataset W_{NCC} starts finding uninformative clusters only after 35 clusters, whereas W_{EH} shows a more stable performance. Intra-cluster distances of clustering on EVMs using c) W_{NCC} and d) W_{DOFF} as compared to clustering of original images using W_{NCC} similarity measure.

clusters with more than 50% uninformative frames are labelled as uninformative. Particularly for this task, W_{EH} yields nicely structured manifolds, where informative and uninformative frames are well separated as shown in Fig.5(a1)-(a3). This is also reflected in the recall-precision curves (Fig.4a-b), where using 7 clusters on this EVM one can cluster apart 70.16% of all uninformative frames (recall) with a precision of 65.61%. Best recall and precision values are summarized in Table 1.

	Video 1			
	Max. Recall	Num. Clusters	Max. Precision	Num. Clusters
W_{NCC}	82.90%	15	74.13%	72
W_{EH}	73.71%	67	88.35%	4

	Video 2			
	Max. Recall	Num. Clusters	Max. Precision	Num. Clusters
W_{NCC}	74.13%	72	90.91 %	40
W_{EH}	71.21%	8	81.25%	30

Table 1. Best recall and precision values of for clustering poor-quality frames.

3.2 Clustering Endoscopic Scenes

The clustering of different segments is performed on optical flow (Sec.2.3a) and intensity based EVMs (Sec.2.3b). Inclusion of temporal constraints for optical flow based EVM requires the use of a larger number of clusters. Therefore, the optical flow and the intensity based EVMs are clustered using 30 and 15 clusters, respectively. The results are compared to K-means clustering perfomed on the original images using the same number of clusters (15 and 30) and W_{NCC} similarity measure. (Fig.5b,c) show the EVMs and examples of the clustered frames. For quantitive evaluation normalized intracluster distances (icd) are measured for all clusters C_i : $\mathrm{icd}(C_i) = \frac{\sum_{x \in C_i} \mathbf{x}-\bar{\mathbf{x}_i}}{\max_{x \in C_i} \mathbf{x}-\bar{\mathbf{x}_i}}$, where $\bar{\mathbf{x}_i}$ denotes the centre of cluster C_i. Fig.4c,d show the decrease in icd when using the W_{DOFF} and W_{NCC} manifolds. This implies that the proposed similarity measures lead to structured manifolds that allow for better separability of the clusters. We further evaluate our results against manual labelling, where contiguous informative frames are labelled to be in the same cluster. The correlation between the ground truth and EVM clusterings is measured by the normalized mutual information, which is

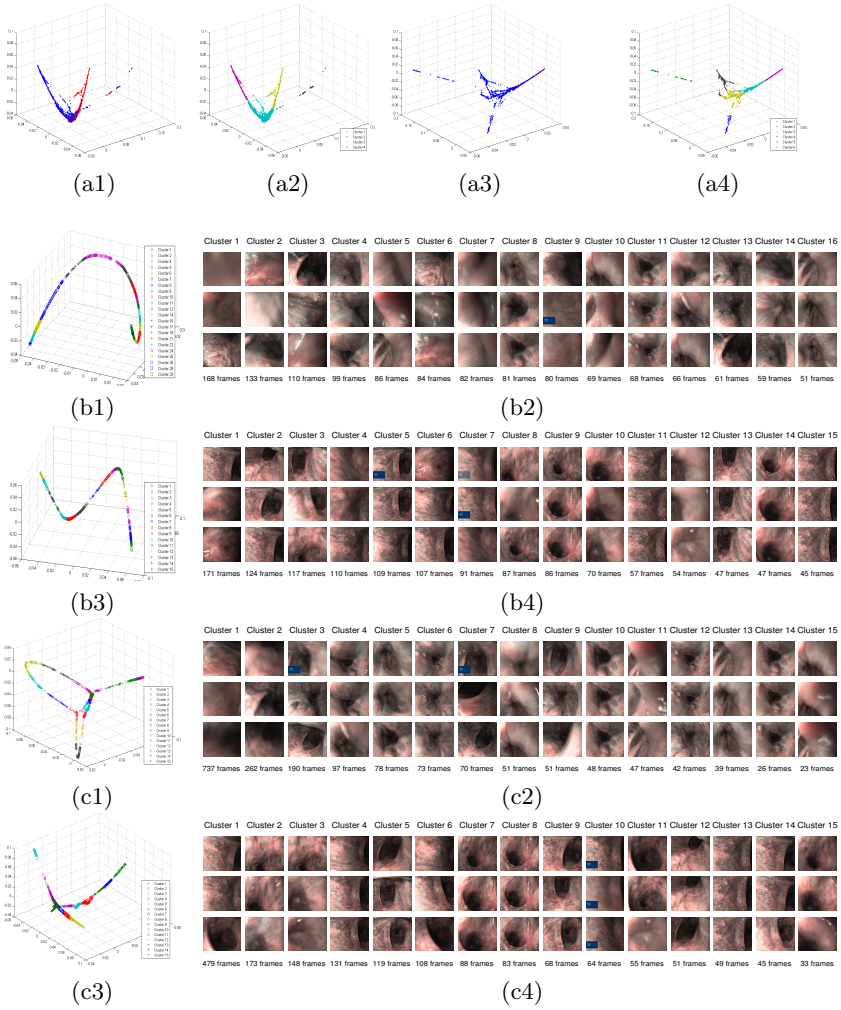


Fig. 5. (a1) and (a3) show the 3-dimensional EVMs of 1. and 2. endoscopic video, respectively. The red points illustrate the poor-quality frames in the ground truth labelling. (a2) and (a4) show the clustering results on the EVMs for the 1. and 2. video, respectively. The use of W_{EH} in manifold learning leads to structured EVMs where the poor-quality frames are clustered together. (b1) and (b3) Largest connected component of 3-dimensional EVM created using W_{DOFF} (Section 2.3) and the clustering on these EVMs for the 1. and 2. dataset, respectively. (b2) and (b4) show 15 example clusters for the 1. and 2. video; each column correspond to one cluster, where the rows show the first, center and the last frames of each cluster, respectively. (c1) and (c3) 3-dimensional EVM created using W_{NCC} (Section 2.3) and the clustering on these EVMs for the 1. and 2. dataset, respectively. (c2) and (c4) show clustering results using 15 clusters for the 1. and 2. video, respectively.

independent of the number of clusters. Clustering on W_{NCC} EVM yields 84.77% and 76.37%. Better results, 87.31% and 75.11%, are obtained with the proposed optical flow based clustering.

4 Conclusion

In this paper, we have proposed an effective framework for clustering endoscopic videos using EVMs. Key technical contribution of the paper includes: 1) we have addressed the task of clustering uninformative frames and endoscopic scenes from a different point of view than the methods in the literature, namely within a generic framework using the inter-frame similarities in an unsupervised manner. Our method provides a compact visualisation of the endoscopic video for subsequent analysis. 2) we have introduced two inter-frame similarity measures for manifold learning, namely rotation invariant energy histograms and divergence of optical flow field. Our experiments demonstrate that the proposed similarity measures yield well structured manifolds and thus lead to accurate clustering. The mathematical framework behind manifold learning has the particular advantage of being extendable by definition of the similarity measures. Therefore, even if particular characteristics of the imaging system changes, EVMs can be easily adopted by changing only the similarity measure.

References

1. Iakovidis, D., Maroulis, D., Karkanis, S.: An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. Computers in Biology and Medicine **36** (2006) 1084–1103
2. Hiremath, P., Dhandra, B., Hegadi, R., Rajput, G.: Abnormality detection in endoscopic images using color segmentation and curvature computation. Neural Information Processing (2004) 834–841
3. Li, P., Chan, K., Krishnan, S.: Learning a multi-size patch-based hybrid kernel machine ensemble for abnormal region detection in colonoscopic images. CVPR **2** (2005) 670
4. Oh, J., Hwang, S., Lee, J., Tavanapong, W., Wong, J., de Groen, P.: Informative frame classification for endoscopy video. Medical Image Analysis **11** (2007)
5. Bashar, M., Kitasaka, T., Suenaga, Y., Mekada, Y., Mori, K.: Automatic Detection of Informative Frames from Wireless Capsule Endoscopy Images. Medical Image Analysis (2010)
6. Iakovidis, D., Tsevas, S., Polydorou, A.: Reduction of capsule endoscopy reading times by unsupervised image mining. Comp. Medical Imaging and Graphics (2009)
7. Balasubramanian, M., Schwartz, E., Tenenbaum, J., de Silva, V., Langford, J.: The isomap algorithm and topological stability. Science **295** (2002) 7
8. Pless, R.: Image spaces and video trajectories: Using isomap to explore video sequences. ICCV (2003) 1433
9. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation **15** (2003) 1373–1396
10. Hartigan, J., Wong, M.: A k-means clustering algorithm. Appl. Stat. **28** (1979)
11. Black, M., Anandan, P.: A framework for the robust estimation of optical flow. ICCV **93** (1993) 231–236