TECHNISCHE UNIVERSITÄT MÜNCHEN
Computer Aided Medical Procedures & Augmented Reality / I16

Endoscopic Video Manifolds for Scene Recognition in Targeted Optical Biopsy

Cana Selen Atasoy

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. D. Burschka
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. N. Navab
	2.	Prof. G-Z. Yang, Ph.D, Imperial College London/UK
	3.	Prof. Dr. R. Pless, Washington University in St. Louis/USA

Die Dissertation wurde am 14.12.2011 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 04.06.2012 angenommen.

Abstract

Recent introduction of a new imaging technology called probe-based confocal laser endomicroscopy (pCLE) enabled real-time and in-vivo visualisation of biological tissue on a microscopic level. The use of pCLE during gastro-intestinal (GI) endoscopy provides the facility for live histological examination of cellular structures without removing any tissue sample. This new technique, referred to as optical biopsy, offers several advantages over the conventional biopsy such as real-time feedback, less-invasiveness, in-vivo and in-situ application. Due to their non-invasive nature, however, optical biopsies do not leave any scar on the tissue and therefore involve the new challenge of re-targeting previous biopsy sites during surveillance examinations.

In this thesis, a novel scene recognition method is presented to support the re-targeting of optical biopsy sites in surveillance GI-endoscopies. Drawing on the mathematical framework of manifold learning, a new representation for endoscopic videos is introduced. In this new representation, the scene recognition problem is reformulated as a clustering and classification task. The low dimensional manifold representation is adapted to facilitate different clustering and classification steps only by changing the notion of similarity between individual endoscopic frames.

Experimental evaluation of the methods is presented on real clinical data. Each step of the proposed framework is evaluated individually on several patient datasets. Final experiments evaluating the complete framework demonstrate the feasibility of the proposed methods as a promising step for assisting the endoscopic expert in re-targeting optical biopsy sites.

Besides addressing the medical problem of optical biopsy re-targeting, this thesis aims at demonstrating new ways to find meaningful representations for very complex datasets such as endoscopic videos. Relying on the tools of spectral theory and their application in manifold learning, the challenge of scene recognition is approached from a novel perspective. From the theoretical point of view, new insights into spectral manifold learning methods are derived using the mathematical equivalence of these techniques with well-studied physical phenomena.

Keywords:
Endoscopy, Optical Biopsy, Manifold Learning, Scene Recognition

Zusammenfassung

Die kürzliche Einführung eines neuen Bildververfahrens, der proben-basierten konfokalen Laser-Endomikroskopie (pCLE), ermöglicht eine Visualisierung von biologischem Weichgewebe auf mikoskopischer Ebene in Echtzeit. Die Anwendung von pCLE während gastro-intestinaler (GI) endoskopischer Untersuchungen bietet die Möglichkeit für sofortige, histologische Untersuchung von zellulären Strukturen ohne Biopsieentnahme. Diese neue Technologie wird als optische Biopsie bezeichnet und bietet viele Vorteile gegenüber klassischer Entnahme-Biopsien, wie beispielsweise das Echtzeit-Information, die minimale Invasion sowie "in-vivo-" und "in-situ-" Anwendung. Zusätzlich hinterlassen die optischen Biopsien aufgrund ihrer Nichtinvasivität keine Narben auf dem Weichgewebe. Andererseits ensteht dadurch eine neue Herausforderung, die optischen Biopsiestellen in Kontrolluntersuchungen wiederzufinden (Re-targeting).

In dieser Dissertation wird eine neue Wiedererkennungsmethode vorgestellt um das Re-targeting von optischen Biopsiestellen bei Kontrolluntersuchungen zu unterstützen. Aufbauend auf den mathematischen Grundlagen des Manifaltigkeit-Lernens wird eine neue Darstellung von endoskopischen Videos eingeführt. In dieser neuen Darstellung ist die Wiedererkennung von endoskopischen Szenen als Clustering- und Klassifizierungsaufgaben umformuliert. Mit dem alleinigen Ändern des Ähnlichkeitsmasses zwischen einzelnen endoskopischen Bildern kann der wenig-dimensionale Manifaltigkeitsraum auf verschiedene Clustering- und Klassifizierungsaufgaben angepasst werden.

Experimentelle Ergebnisse der Methoden sind anhand von klinischen Datensätzen präsentiert. Jeder Schritt des vorgestellten System ist auf mehreren Patientendatensätzen ausgewertet. Gesamtergebnisse, die das komplette System evaluieren, zeigen die Durchführbarkeit der präsentierten Methoden um den Endoskopie-Experten beim Re-targeting von optischen Biopsiestellen zu unterstützen.

Neben dem Ansatz für das medizinische Problem von optischen Biopsie-Targeting, soll diese Dissertation auch neue Wege aufzeigen um sinnvolle Darstellungen von komplexen Datensätzen wie endoskopische Videos zu finden. Unter Berufung auf die Spektraltheorie und ihre Anwendung im Manifaltigkeit-Lernen wird die Herausforderung der Wiedererkennung von endoskopischen Szenen aus einer neuen Sichtweise betrachtet. Unter Verwendung der mathematischen Äquivalenz zwischen den Methoden des Manifaltigkeit-Lernens und den physikalischen Phänomenen werden, ausgehend von einer theoretischen Perspektive, neue Einblicke in diese Techniken hergeleitet.

Schlagwörter:
Gastrointestinale Endoskopie, Optische Biopsie, Manifold Learning,
Szenen-Wiedererkennung

To my father, my greatest support

Acknowledgments

I would like to express my deepest appreciation to my two supervisors, Professor Nassir Navab and Professor Guang-Zhong Yang, for giving me the privilege of working with two great minds. When starting this Ph.D. project, I was advised by many that this kind of strong collaboration with monthly travel will be very difficult and not long lasting. Looking back now, I see that Professor Navab and Professor Yang not only made this journey possible but also supported me constantly with their great motivation and extra-ordinary supervision. Thank you for all your guidance and scientific feedback, which I received not only in our regular meetings but also in several discussions via skype, sometimes late hours or on the weekends, and especially thank you for your encouraging me in following my ideas.

I owe special thanks to Diana Mateus for countless discussions, for late deadline nights and also for her friendship. Without your amazing support and willingness to follow crazy ideas, this work would not have been possible! I also would like to thank Andreas Georgiou for sharing his enormous knowledge in physics, for all the inspiring and fun coffee meetings. I have learned so much from you! I am deeply grateful to Professor Alexander Meining for his medical supervision and his great motivation for this collaboration. I also thank Tobias Lasser for many tough but fruitful discussions.

I have had the great luck of being welcomed in two amazing research groups and owe many thanks to my colleagues in both labs, who made this experience so valuable and enjoyable. Special thanks to Matina, Pete, Valentina, Ben, Darko, Olivier, Keili, Rob, Dan, Steffi, Alessio and Dave Noonan for creating such fun working environments. Thanks to Julien, Rach, Andy, Doug, Alex, Chris, Jim, Johannes, Dan Elson, George, Ka-Wai and all my colleagues from Imperial College for the regular Thursday meetings. I thank Stefan Hintersteusser, Jose, Max, Tobias Reichl, Martin Groher, Hauke, Loren, Athanasios, Martin Horn, Nicolas, Cedric and Mehmet for the fun coffee and lunch breaks at CAMP. Many thanks to Karim, Wolfi, Christian, Marco V., Toby Wood, Marco Feuerstein and Su-lin for all the nice conference trips. Thanks to Lichau for the morning chats and taking special care of my plant. I owe many thanks to Julien and Slobodan also for the funniest French lessons. Besides, I would like to thank all my friends in Munich and London for making me feel home in both cities.

I would like to express my gratitude to Martina Hilla for supporting me in any administrative matter at any time! I also thank Raphaele Raupp for all her organisational help. I wish to thank TUM Graduate School for supporting my Ph.D. Project. I am also thankful to Professor Tomaso Poggio for his invitation and his warm welcome during my short visit.

Finally, I thank my parents, Seda and Cihad Atasoy and my sister Sıla, for their greatest support along every step of this amazing journey!

Contents

Part I

Introduction

Chapter 1

Medical Application

'Diagnosis is not the end, but the beginning of practice.'

Martin H. Fischer

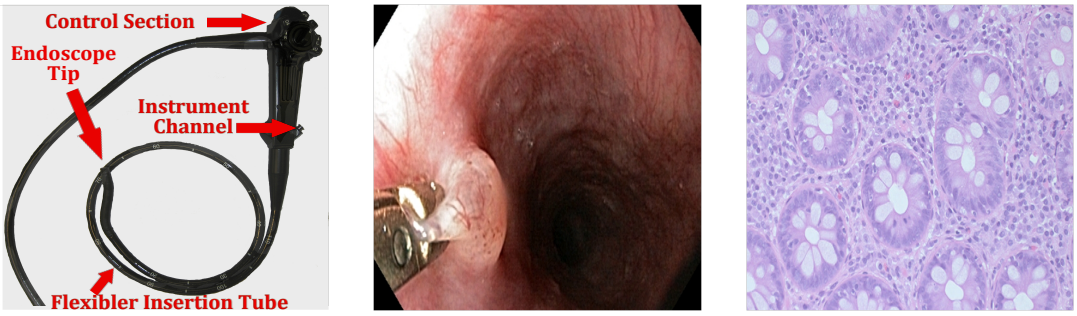
Gastrointestinal (GI) endoscopy is a widely used clinical technique for visualising the digestive tract. Currently, diagnosis and surveillance[1] of several diseases of the GI tract are performed via an endoscopic examination. Among other applications, GI endoscopy is critically important for the early diagnosis and surveillance of the highly lethal oesophageal cancer, called oesophageal adenocarcinoma (OAC). OAC is the most rapidly increasing cancer in the United States and the Western World [Fitzgerald, 2004, Sharma et al., 2006]. Despite the treatment options including surgery, endoscopic mucosal resection and photodynamic therapy, the 5-year mortality of this cancer is more than 80% [Fitzgerald, 2004], and around 90% in most Western populations [Wani and Sharma, 2006]. The primary reason of this low survival rate in OAC is the advanced stage at diagnosis [Sharma et al., 2006].

Barrett's Oesophagus (BO), referring to an abnormal change of the oesophageal mucosa caused by gastro-oesophageal reflux, has attracted increasing interest because of its strong association with OAC. BO is the only recognized precursor of OAC and the risk of developing oesophagus cancer is 30-40-fold higher for patients with BO compared to the general population [Sharma et al., 2006]. Surveillance endoscopies performed at regular intervals can allow for the diagnosis of this highly lateral cancer at an early stage and prevent its progression into invasive cancer. Therefore, periodic surveillance examinations by GI endoscopy together with systematic biopsy are crucial for patients diagnosed with BO or OAC.

[1] Surveillance is described as the periodic testing of individual patients known to be at high risk for disease [Wani and Sharma, 2006].

1.1 Gastro-Intestinal Endoscopies

Currently, the gold standard for the surveillance protocol consists of regularly performed GI endoscopies. During these procedures a flexible endoscope is inserted through the mouth of the patient and guided through the oesophagus to the z-line[2]. The oesophageal tissue is examined under endoscopic guidance and biopsies are acquired from suspicious tissue regions as well as every 1-2 cm along the oesophagus. To this end, a small sample of the tissue is removed for histopathological examination (Figure 1.1). For surveillance examinations it is highly important to acquire biopsies from the same locations as in the diagnostic endoscopy for tracking the progress of the disease and comparison. This procedure, known as Seattle protocol, requires a large number of biopsies, and is very time consuming, invasive and costly [Egger et al., 2003]. Furthermore, it incurs additional risks due to the large number of biopsy specimen and consequently lengthy healing period of the affected tissue [Egger et al., 2003]. Any potential to increase the cost-effectiveness and to reduce the invasiveness of this protocol would improve the current surveillance procedure.



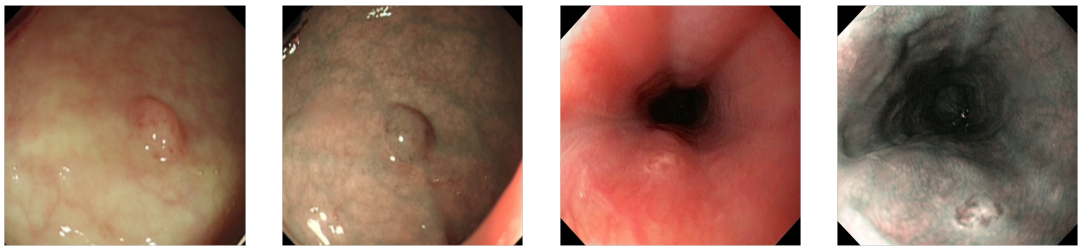
(a) (b) (c)

Figure 1.1: Gastro-Intestinal endoscopies. (a) State-of-the art flexible endoscope used in Gastro-Intestinal endoscopy. (b) Endoscopic view of the conventional biopsy acquisition by removing a tissue sample. (c) Histopathology of the conventional biopsy. Figure (c) is reprinted from [Meining et al., 2007].

1.1.1 White Light Endoscopy (WLE)

A standard flexible endoscope consists of three main parts; a control section, a flexible insertion tube and a connector section. The hand held control section (Figure 1.1(a)) conduces to the manipulation of the endoscope tip via horizontal and vertical bending motion. Furthermore, the control section has an entry port to the instrument channel through which other tools can be inserted (Figure 1.1(a)). The flexible insertion tube contains a charge-coupled device (CCD) for colour image generation, a light guide

[2]Z-line refers to the gastro-oesophageal junction between the oesophagus and the stomach.



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

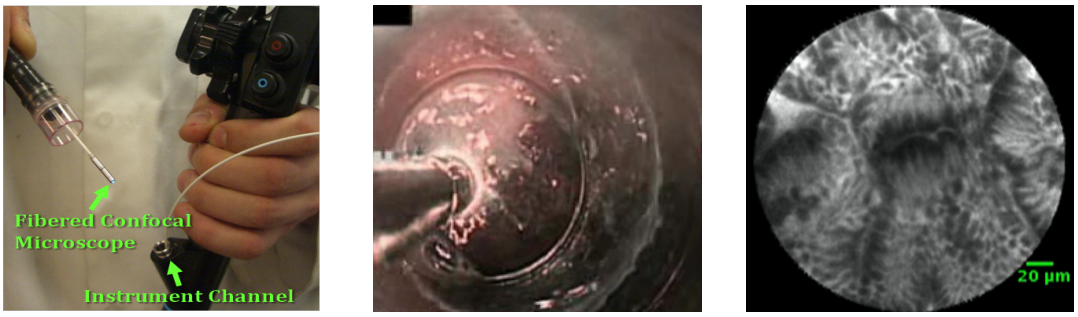
Figure 1.2: Comparison between white light endoscopy (WLE) and narrow-band endoscopic imaging (NBI). (a) and (c) show endoscopic frames acquired with WLE. (b) and (d) illustrate the same anatomical locations as in (a) and (c), respectively, captured with NBI.

illumination system and an objective lens, all attached at the endoscope tip. The endoscope is connected to an image processor and a light and electric source through the connector section. During acquisition, the tissue is illuminated with white light and the reflected red, green and blue images are sent through a colour CCD attached at the endoscope tip and finally transmitted to the image processor.

This imaging technology using a normal white light illumination, referred to as white light endoscopy (WLE), is the conventional imaging technology utilized in GI endoscopic examinations. Nowadays, state-of-the art GI endoscopes are further equipped with a more advanced imaging technique called, Narrow Band Endoscopic Imaging (NBI). During the acquisition, the endoscopic expert can simply switch between the two endoscopic imaging modalities.

1.1.2 Narrow-Band Endoscopic Imaging (NBI)

NBI is a novel imaging technique which combines the conventional endoscopy with an optical image enhancement technology. By the use of narrow-bandwidth filters conventional white light imaging is narrowed to selected wavelengths of the spectrum. Interaction between the light with different wavelengths and the mucosa results in different images at distinct levels and increases the contrast between the epithelial surface and the subjacent vascular pattern [Gheorghe, 2006] (Figure 1.2). Thus, this technique allows for imaging of superficial tissue structures and emphasizes surface features, such as vessels and mucosal patterns. NBI is mostly combined with a magnifying endoscope, which enables precise examination of colour and structure variations of the mucosa. Therefore, NBI is considered to be a useful supporting method to observe the endoscopic findings of early cancer by providing enhanced anatomical and structural information of the mucosa [Gono et al., 2004, Nonaka et al., 2006]. However, a clear determination of the malignant region is still difficult and the diagnosis is based on the interpretation of the observed anatomy by the physician.



(a) (b) (c)

Figure 1.3: Probe-based Confocal Endomicroscopy (pCLE). (a) pCLE inserted through the instrument channel of the flexible endoscope. (b) pCLE as viewed from the endoscope while acquiring optical biopsy during GI endoscopy. (c) An example optical biopsy acquired *in-vivo* using pCLE during a GI-endoscopic procedure.

In the literature, some studies evaluating the performance of NBI have reported significant improvement in relation to WLE [Su et al., 2006, Tischendorf et al., 2007, Rastogi et al., 2009], whereas other studies did not find any significant difference in the performance of these the two imaging techniques [Rex and Helbig, 2007, Kaltenbach et al., 2008, Paggi et al., 2009]. Using eye-tracking as a means to evaluate human perception [Meining et al., 2010], we also presented a comparative study of NBI and WLE. Besides its suitability for the perception of human observers, endoscopic frames acquired by NBI contain more information compared to the conventional WLE due to the visibility of submucosal structures and vascular patterns. The enhancement of the vascular pattern and tissue structures in NBI increases the information captured in the image content. Therefore, the experimental studies presented in the rest of this thesis are performed on NBI endoscopic videos. For simplicity of the presentation, the used NBI datasets will be simply referred to as endoscopic videos.

1.1.3 Probe-based Confocal Laser Endomicroscopy

Recently, a new technology called *probe-based confocal laser endomicroscopy (pCLE)* became available, which allows for *in-vivo* visualisation of the tissue at a cellular level. A fibered confocal micro-probe is inserted through the instrument channel of a state-of-the-art endoscope (Figure 1.3(a)). In contact with the tissue (Figure 1.3(b)), pCLE visualizes the tissue at a microscopic scale allowing for *"optical biopsies"* (Figure 1.3(c)). Due to its non-invasive nature and the real-time, in-vivo feedback, pCLE provides significant advantages over the conventional biopsy. High agreement between the optical biopsy and conventional histopathology results suggest that pCLE will be increasingly used in daily clinical routine [Meining et al., 2007, Bajbouj et al., 2010].

Besides all 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 [André et al., 2009a, André et al., 2010b, André et al., 2009b], 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 [André et al., 2010a]. A further challenge introduced by pCLE is the re-targeting of previous optical biopsy sites. In contrast to the conventional endoscopy, the non-invasive 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.

1.2 Re-targeting of Optical Biopsy Sites

With the development of the pCLE, the newly introduced task of optical biopsy re-targeting has drawn attention in the medical imaging community. A complete solution for assisting the endoscopic expert in re-targeting the optical biopsy sites in serial endoscopic examination involves two challenges.

Firstly, the relevant frames of the surveillance endoscopic video showing a previous optical biopsy location should be retrieved during the surveillance examination. This provides *intra-video localization* of the optical biopsies; i.e. extraction of video segments containing an optical biopsy site. Once an optical biopsy scene is recognized during the surveillance endoscopy, an accurate localization of the confocal microprobe within the endoscopic frame can be performed. We refer to this second challenge as *intra-frame localisation* of the optical biopsy locations.

1.2.1 Intra-frame Localization

In the last years, several approaches have been proposed for point-based re-localization of the fibered confocal microprobe [Allain et al., 2009, Allain et al., 2010, Mountney et al., 2009]. Allain *et al.* present a method for probe re-localization based on the epipolar geometry between endoscopic frames [Allain et al., 2009, Allain et al., 2010]. Mountney *et al.* introduce the simultaneous localisation and mapping (SLAM) framework in order to create a 3D model of the tissue surface containing the optical biopsy sites. The presented method provides an augmented view of the endoscopic scene together with the point-based optical biopsy mappings in order to support the re-localisation of the optical probe [Mountney et al., 2009]. In Chapter 9 of this thesis, we also present a deformable wide baseline matching method which was previously explored in [Atasoy et al., 2009]. Our proposed method allows for establishing point correspondences between two frames showing the same scene in the diagnostic and surveillance endoscopy videos.

The crucial first step for the application of these intra-frame localization methods, however, is the intra-video localisation. Any of the developed intra-frame localisation approaches necessitates the retrieval of the relevant endoscopic frames in real-time during

the surveillance endoscopy and calls for the development of intra-video localisation methods.

1.2.2 Intra-video Localization

In terms of medical contributions, the main focus of this thesis lies in the development of an intra-video localisation framework. To address this problem, a clustering and classification method is introduced. First, patient specific endoscopic segments are defined based on a two step clustering of the diagnostic endoscopy video [Atasoy et al., 2010b, Atasoy et al., 2012] as presented in Chapter 7. Two different approaches for the classification of surveillance endoscopic frames are investigated. The first method, presented in [Atasoy et al., 2012], focuses on the classification of surveillance endoscopic frames individually and is explained in detail in Section 8.1. The second approach is designed to establish scene correspondences between diagnostic and surveillance videos, even in the presence of severe structural changes of the oesophageal tissue caused by the treatment of the patient. Relying on cluster correspondences acquired with user interaction, this method extends the individual clustering method to a video-based scene matching [Atasoy et al., 2011] as presented in Section 8.2 in this thesis.

Chapter 2

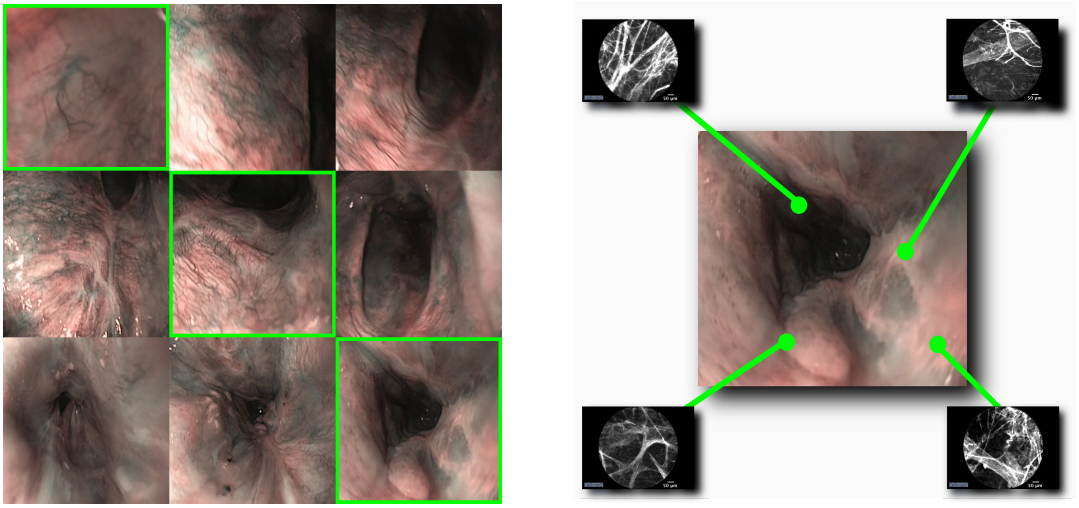
Analysis and Problem Statement

> 'If we knew what it was we were doing, it would not be called research, would it?'
>
> Albert Einstein

From the medical point of view, the aim of this thesis is to introduce a system for assisting the endoscopic expert in re-targeting the optical biopsy locations during surveillance examinations. As discussed in Chapter 1, the re-targeting task involves two steps; i.e. recognition of the frames containing an optical biopsy site (*intra-video localisation*) and aiding the expert for guiding the micro-probe to the exact location of the optical biopsy within the retrieved frames (*intra-frame localisation*). The main focus of the work presented in this thesis lies in providing a scene recognition method for the intra-video localisation task. In this chapter, we discuss the precise steps involved in such a scene recognition system from the image processing perspective.

2.1 Data Analysis

After the first screening endoscopy, a patient diagnosed with BO or OAC is recommended to undergo periodic surveillance examinations in 3 to 4 months intervals. If a diagnosis is established in the screening examination, we refer to the acquired dataset as *diagnostic endoscopy*. According to the current clinical routine, the complete video data acquired during the diagnostic endoscopy is provided as input for the re-targeting framework. Furthermore, we can also rely on a marking of the optical biopsy locations on the frames of the diagnostic endoscopy. Figure 2.1 illustrates sample frames of a diagnostic endoscopy and a marking of optical biopsy locations. The long time interval of 3 to 4 months until the next surveillance examination allows for detailed offline processing



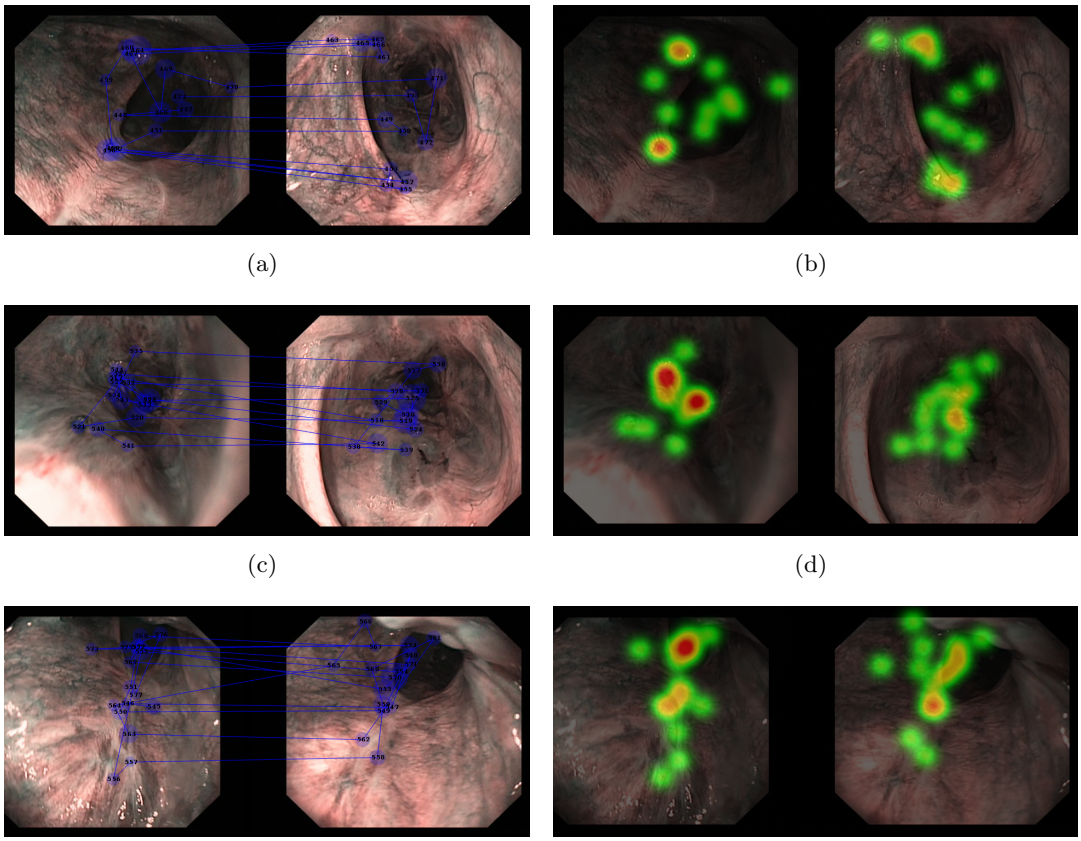
(a) (b)

Figure 2.1: Data acquired during diagnostic endoscopy. (a) Example frames showing different scenes of the diagnostic examination, where the frames containing an optical biopsy site are marked with green. (b) Mapping of the optical biopsy sites into the endoscopic view. It is supposed that the marking of optical biopsy frames and exact locations within the frame are given as the input for the re-targeting framework. Intra-video localisation during surveillance endoscopy requires the recognition/retrieval of frames showing the same anatomy as marked in (a), whereas intra-frame localisation assists the endoscopic expert in the exact localisation of the optical micro-probe at the marked optical biopsy sites in (b).

of the diagnostic endoscopy, but requires an efficient online scene recognition that can be applied in real-time during a surveillance endoscopy. For in-depth analysis of the endoscopic data, which we are confronted with, we first explore the perception of the endoscopic expert for such endoscopic datasets. The goal behind this exploration is to gain insights into the visual cues and patterns used by an endoscopic expert before developing the precise steps of our re-targeting framework.

2.2 Analysis of Expert's Perception

Given enough time for analysis, an endoscopic expert can recognize the corresponding scenes between two examinations, even in the presence of severe structural changes as illustrated in Figure 2.4. Such significant changes in the visual appearance of the tissue can occur due to the treatment of the patient; e.g. as when the patient undergoes chemo-therapy or endoscopic mucosal resection. During the endoscopy, an additional task for the expert such as the visual analysis of the scenes for recognition is not suitable due to the already demanding nature of an examination and the time constraints.



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

Figure 2.2: Fixations of the endoscopic expert while evaluating endoscopic scenes. (a), (c) and (e) show the locations of the fixations, where their temporal order is marked using numbering. (b), (d) and (f) display durations and locations of the fixations where warmer colours indicate longer durations.

In order to analyse expert's perception on endoscopic datasets, we perform an eye-tracking study. To this end, the endoscopic expert is shown several frame pairs displaying corresponding and non-corresponding scenes. The expert is asked to evaluate whether the frames show the same anatomical region. No time constraint is applied for the evaluation; i.e. the visual stimuli is changed only after the expert reported his decision. Using an eye-tracking system (Tobii 1750 eye tracker with integrated 17" monitor), the eye-movements of the expert are recorded while examining the endoscopic frame pairs. Location and duration of the fixations[1] are extracted. Figures 2.2(a), 2.2(c) and 2.2(e) show the fixation points of the expert by taking into account their temporal order. Durations of the fixations are illustrated using a heat map in Figures 2.2(b), 2.2(d) and 2.2(f).

[1] A fixation is defined if the gaze duration at the same location was longer than 100 ms.

This study, which initiated for the analysis of expert's perception, is continued by exploring several properties seen in human perception. Inspired by some properties pointed out by the Gestalt theory [Koffka, 1999], we presented a mathematical model based on wave interference [Atasoy et al., 2010a]. It is demonstrated that this model exhibits similar properties studied by the Gestalt theory for human perception. The introduced mathematical model and its application for pattern recognition can be found in Appendix A. Furthermore, the same experimental set-up designed for these eye-tracking experiments is utilized for another evaluation study comparing the perception of NBI to conventional WLE [Meining et al., 2009, Meining et al., 2010].

Based on the eye-tracking study presented in this section, we can conclude that experts utilize prior knowledge in extracting information from the endoscopic datasets. We did not observe a regular pattern extraction such as textured regions due to high vasculature or structural landmarks performed by the endoscopic expert while perceiving these datasets.

Inspired by the outcome of this study, we introduce a new representation for endoscopic videos which represents each frame in relation to other frames of the video instead of extracting particular patterns/features from each frame. This new representation, called *"Endoscopic Video Manifold" (EVM)*, is created by learning the underlying non-linear manifold of an endoscopic video. Thus, the EVM representation respects complex non-linear relations within a dataset while representing each endoscopic frame in a low dimensional space. In the re-targeting framework presented in this thesis, all processing of endoscopic videos is performed in this new EVM representation, as explained in detail in Part II.

2.3 Challenges

In order to assist the endoscopic expert in optical biopsy re-targeting, we introduce a framework based on clustering of diagnostic endoscopy video and classification of surveillance endoscopy frames. In this section, we discuss the challenges involved in clustering and classification tasks when performed on endoscopic datasets.

2.3.1 Uninformative Frames

Particular conditions of 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 (Figure 2.3(e)) or out of focus imaging of the endoscope (Figure 2.3(f)), frames filled with bubbles caused by the fluid inside the oesophagus (Figure 2.3(g)) and frames with large specular reflections (Figure 2.3(h)). Presence of these uninformative frames leads not only to poor accuracy in the clustering and classification tasks but also complicates the post-processing of the video by the endoscopic expert which is required for several endoscopic applications.

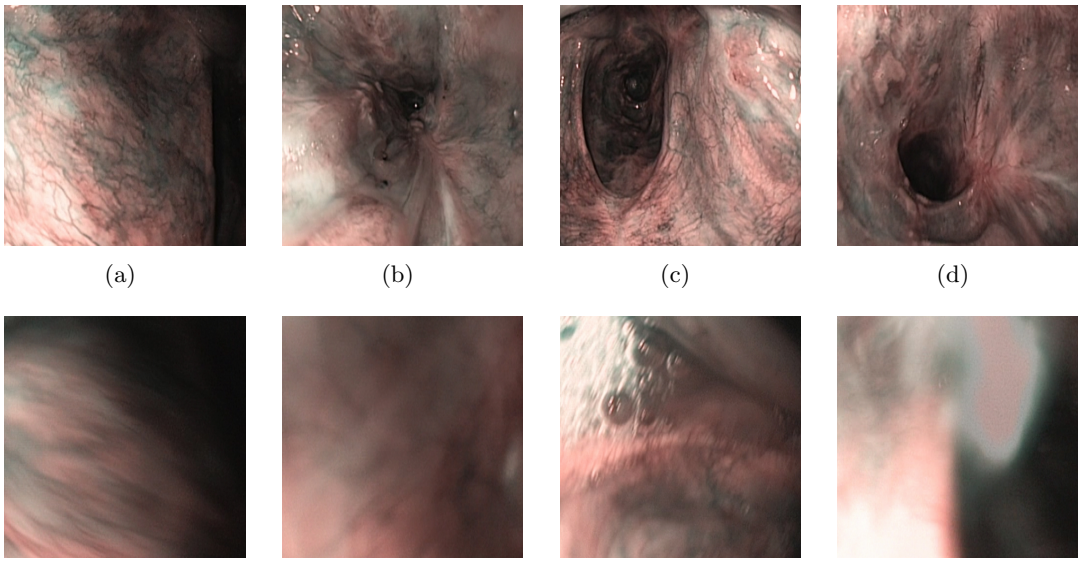


Figure 2.3: Several challenges encountered in endoscopic videos. (a), (b), (c) and (d) show examples for informative frames acquired by a state-of-the art GI endoscope. (e), (f), (g) and (h) illustrate examples for uninformative frames caused by (e) motion blur, (f) out-of-focus imaging, (g) bubbles due to the liquid inside the oesophagus and (h) specular highlights.

In the literature, detection of uninformative frames has been studied for different endoscopic procedures, such as capsule endoscopy [Bashar et al., 2008, Bashar et al., 2010, Iakovidis et al., 2010] and colonoscopy [Arnold et al., 2009, Oh et al., 2007]. The main focus of these studies is on defining specific features such as colour or texture in order to detect uninformative frames within an endoscopic video. The endoscopic expert is then provided the remaining informative frames instead of the whole content for post-processing.

In our particular application, the definition of meaningful endoscopic segments requires elimination of the uninformative frames without losing relevant information. To address this challenge, we *cluster* the informative and uninformative frames of GI endoscopic videos in an *unsupervised* manner and allow the *expert* to easily select the frames (clusters) for further processing. This step is explained in detail in Chapter 7. This labelling and elimination of uninformative frames prior to further processing leads to meaningful endoscopic segments, which are desired for the online classification of the surveillance endoscopic frames.

2.3.2 Endoscope and Tissue Motion

Scene recognition in endoscopic videos is a challenging task not only due to the camera viewpoint change (Figures 2.4(c), 2.4(d)), which is a well-studied problem in compute vision, but also because of the deformation of the tissue (Figures 2.4(e), 2.4(f)) and prevalence of homogeneous tissue regions (Figures 2.4(g), 2.4(h)). Furthermore, as the endoscope is very close to the tissue, small differences in the visible scales of the same feature can cause a significant change in the visual content (Figures 2.4(f), 2.4(h)).

The issue of view-invariance is an extensively studied problem in the computer vision community[2]. One of the most commonly used approach for this problem is the bag-of-features model where an image is represented as a set of features without taking their locations into account. First, a feature detection is performed in order to extract distinctive features from the image content. This step yields a set of local image patches. Afterwards, a normalization operator is applied to each patch in order to achieve invariance to the desired transformations such as scale and rotation or affine transformations. Each normalized patch is represented with a descriptor vector. Extensions of this model for scene description is achieved by creating a visual vocabulary by clustering the descriptor vectors and representing each image as a histogram of these visual words. This approach has been originally developed for object and scene recognition in computer vision applications [Sivic and Zisserman, 2009]. In medical imaging community, it has also been extended for retrieval of endomicroscopic images [André et al., 2009a, André et al., 2010b] and videos [André et al., 2011] as well as for tumour classification in magnifying NBI images [Tamaki et al., 2010].

In Chapter 9, we present a region matching method, based on our previous study [Atasoy et al., 2009], which relies on view-point invariant feature extraction to support the endoscopic expert in intra-frame localisation [Atasoy et al., 2009]. Instead of using a bag-of-features approach, however, our method involves a Markov random field (MRF) model. The MRF model is designed to take into account the location information while remaining invariant to large degree of view-point change and tissue deformations. For the intra-video localisation, however, the application of such feature based approaches on endoscopic datasets is very challenging. This is mainly due to the lack of distinctive image features in most of the endoscopic scenes as shown in Figures 2.4(g) and 2.4(h). Therefore, the intra-video approach presented in this thesis relies on the introduced EVM representations. Thereby, the relations between individual endoscopic frames are considered in a manner that intrinsically exhibits large degree of invariance to endoscope viewpoint change and deformations of the oesophageal tissue. A discussion on this intrinsic invariance observed in EVMs is provided in Section 6.7.2.

[2] A comprehensive review of the presented techniques is beyond the scope of this thesis. As an example, we discuss one of the commonly used techniques, which also relates to our intra-frame localisation presented in Chapter 9

2.3.3 Structural Changes of the Oesophageal Tissue

Besides the common challenge of view-invariance, scene recognition in endoscopic videos is further complicated by structural changes of the oesophageal tissue. In the case of a treatment such as chemo-therapy or mucosal resection, visual appearance of the same anatomical region can change significantly. Figures 2.4(e) and 2.4(f) illustrate examples for the appearance change of the same anatomical region shown in two different examinations, where the patient underwent chemo-therapy.

As discussed in Section 2.2, an endoscopic expert can successfully recognize such endoscopic scenes. Given the challenging circumstances of an endoscopic examination, however, such visual assessment during the actual procedure is not suitable for the clinical routine. In order to provide an online scene recognition even in these challenging cases, we extend our clustering and classification method by inter-examination scene matching. This extension involves acquisition of an additional endoscopic video prior to the actual surveillance examination and expert's feedback in establishing scene correspondences *prior to the actual examination*. This pre-interventionally acquired video, together with expert's feedback creates the link between the corresponding scenes of the diagnostic and surveillance endoscopies even in the presence of severe structural changes. This way, very challenging inter-examination re-targeting is reformulated into the plausible problem of intra-examination frame recognition as explained in detail in Section 8.2.

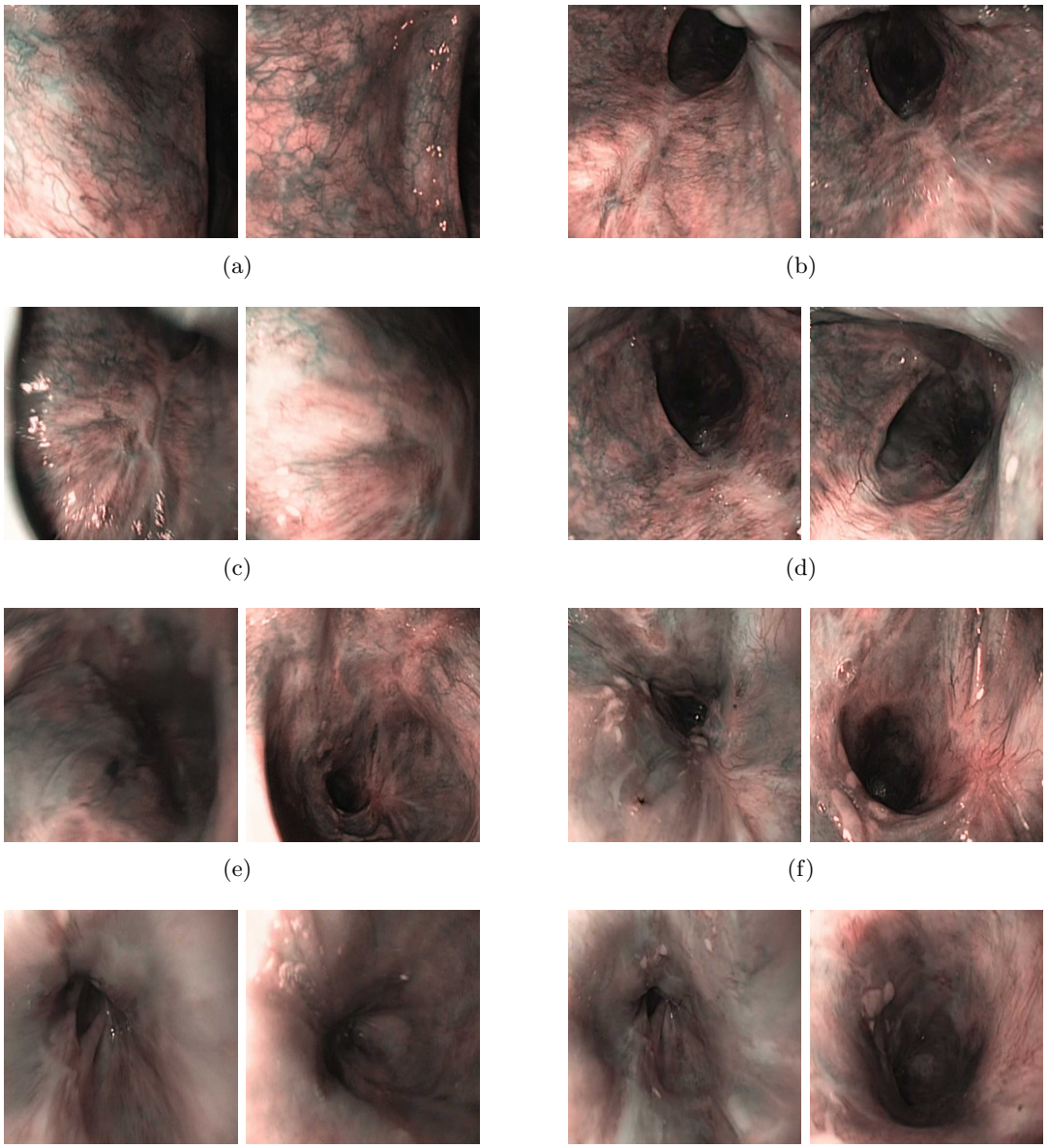


Figure 2.4: Corresponding endoscopic scenes from diagnostic and surveillance endoscopies. Two scenes shown in two different examinations (a), (b) from the similar endoscopic viewpoints, (c), (d) from different viewpoints of the endoscope. (e) and (f) illustrate scenes with severe changes in the visual appearance of the same anatomical region. Such structural changes can occur due to the treatment. In this particular case the patient underwent chemo-therapy between the examinations, where the scenes before and after the therapy are illustrated on the left and he right hand sides, respectively. (g) and (h) illustrate scenes, which lack distinctive image features, from two different examinations.

Chapter 3

Contributions

> 'I must admit that I personally measure success in terms of the contributions an individual makes to her or his fellow human beings.'
>
> Margaret Mead

This thesis presents a novel scene recognition method to assist the endoscopic expert in re-targeting optical biopsy sites in surveillance GI-endoscopies. To address this medical problem, two different challenges are identified; i.e. intra-video and intra-frame localization of the optical biopsy sites. The first challenge involves the recognition of the endoscopic frames showing a previous optical biopsy location during a surveillance examination, whereas the second challenge requires an accurate localisation of the optical biopsy sites within the recognized frame. In this study, the main focus is directed towards the intra-video localization (Chapters 7 and 8) while a complementary inter-frame localisation method is presented in Chapter 9.

In terms of medical contributions, a novel clustering and classification framework is developed to address the intra-video localisation task. The proposed framework is based on the offline (post-procedural) processing of the endoscopic video acquired during the diagnostic endoscopy in order to define *patient specific endoscopic segments (PSESs)*. The second stage involves online (intra-procedural) classification of new endoscopic frames acquired during the *surveillance endoscopy* as belonging to one of these pre-defined segments. The individual frame classification is further extended by inter-examination scene correspondences in order to handle significant structural changes of the oesophageal tissue caused by the treatment techniques such as chemo-therapy.

The technical contributions of this thesis are two-fold. First, a new representation for endoscopic videos, endoscopic video manifold (EVM), is introduced to facilitate the involved clustering and classification steps. The low dimensional EVMs respect non-linear relation between individual frames and provide meaningful representations for

very complex datasets such as GI endoscopic videos. As a second technical contribution, simple ways to adapt the EVM representation to each of the addressed clustering and classification tasks are demonstrated. Taking advantage of the particular mathematical framework of manifold learning, this is achieved only by changing the notion of similarity between individual frames.

From a theoretical point of view, this thesis aims at presenting new insights into spectral manifold learning methods. In particular, an interpretation for the Laplacian eigenmaps method is discussed by drawing on its mathematical equivalence with the well-studied physical phenomenon of stationary waves.

The reminder of this thesis is subdivided into two different parts. The first part (Part II) contains the theoretical background and the technical details of the proposed EVM representation, whereas its application for optical biopsy re-targeting is introduced in the second part (Part III) .

Part II: ENDOSCOPIC VIDEO MANIFOLDS

CHAPTER 4: MANIFOLD LEARNING

This chapter provides an introduction into non-linear manifold learning and gives an overview of the spectral manifold learning methods within a common framework.

CHAPTER 5: THEORETICAL INSIGHTS

This chapter introduces the necessary background for the spectral manifold learning. Recalling the relations between spectral theorem and repreoducing kernel Hilbert spaces, the problem solved in the spectral manifold learning methods is reformulated. For the particular manifold learning method, Laplacian eigenmaps, which is used to create EVMs, further theoretical insights are gained by drawing on its mathematical equivalence with the well-studies physical phenonemon of stationary waves.

CHAPTER 6: CREATING ENDOSCOPIC VIDEO MANIFOLDS

This chapter presents the methodological basis of the introduced EVM representation based on our previous explorations in [Atasoy et al., 2010b, Atasoy et al., 2011, Atasoy et al., 2012]. The individual steps for creating EVMs from a GI endoscopic video are introduced. An extension for including temporal constraints into the EVM representation is also presented.

Part III: TARGETED OPTICAL BIOPSIES

CHAPTER 7: Clustering Diagnostic Endoscopy

Application of the EVM representation for clustering the diagnostic endoscopic video is presented in this chapter. Two different clustering tasks are addressed. The diagnostic video is first clustered into informative and uninformative groups. In the second task, patient specific endoscopic segments are created by clustering the diagnostic endoscopy into different visual segments. Quantitative evaluation for both clustering steps is performed based on the experimental set-up used in [Atasoy et al., 2010b, Atasoy et al., 2012]. For the second clustering step, the efficiency of several EVM representations is demonstrated in comparison to original image representation and principal component analysis.

CHAPTER 8: SCENE RECOGNITION IN SURVEILLANCE ENDOSCOPY

In the second stage of the proposed framework, to each frame of the surveillance endoscopy a patient specific endoscopic segment is assigned via classification. To this end, two different classification approaches are introduced; first, classification of individual endoscopic frames, and second, classification using inter-examination scene correspondences.

The first approach, previously explored in study [Atasoy et al., 2012], involves the projection of new endoscopic frames into the low dimensional representation and a nearest neighbour classification in this low dimensional EVM representation.

The second approach, first presented in [Atasoy et al., 2011], provides an extension of the individual frame classification and relies on two run surveillance endoscopies. By introducing an additional endoscopic dataset acquired prior to the surveillance examination, the presented method creates a link between the scenes of the diagnostic and surveillance examinations. This extension allows us to address the inter-video localisation task even in the presence of significant structural changes caused by the treatment of the patient with techniques such as chemo-therapy or mucosal resection.

CHAPTER 10: INTER-FRAME LOCALISATION

This chapter presents a method for deformable wide-baseline matching between two endoscopic frames in order to support the intra-frame localisation of the optical biopsy sites within the retrieved surveillance endoscopic frames. The presented approach is based on our explorations in [Atasoy et al., 2009] and models task of matching local image regions within the Markov random field framework and future directions.

CHAPTER 11: CONCLUSIONS

This chapter presents a short summary of the presented methods and contributions of this thesis, together with a brief discussion on potential improvements.

Appendix

A. A PERCEPTUALLY INSPIRED PATTERN DESCRIPTOR

Analysis of the expert's knowledge and perception on endoscopic images as discussed in Chapter 2 resulted in the study of the basic properties seen in human perception. This chapter provides a brief overview of these properties and introduces a mathematical model where similar properties can be observed. A perceptually inspired pattern description method developed in our earlier study [Atasoy et al., 2010a] is explained and its application for several pattern recognition tasks are demonstrated.

B. ABBREVIATIONS

This chapter includes a list of abbreviations used throughout this thesis.

C. PUBLICATIONS

All publications contributed to the scientific community during this work are listed in this chapter.

Part II

Endoscopic Video Manifolds (EVMs)

Chapter 4

Manifold Learning

'It requires a very unusual mind to undertake the analysis of the obvious.'

Alfred North Whitehead

Finding meaningful representations for large datasets is an essential step in solving several medical image processing tasks. When working on a large set of images, such as a video sequence, representing each image as a collection of pixel intensities may not be very suitable due to the curse of dimensionality [Beyer et al., 1999]. Seeking more suitable representations for the high dimensional data, non-linear manifold learning methods have been first introduced as a tool for dimensionality reduction [Tenenbaum et al., 2000, Roweis and Saul, 2000]. The task of reducing the dimensionality was formulated as finding meaningful low dimensional structures hidden in high dimensional observations [Tenenbaum et al., 2000]. A potential connection between the human perception of visual stimuli and image manifolds have also been discussed in [Seung and Lee, 2000].

In a medical setting, although acquired datasets are typically of high dimension, the imaged anatomy varies smoothly and in a non-linear fashion within the dataset; e.g. slices of magnetic resonance imaging (MRI) or computed-tomography (CT) volume, observations of a deforming object under breathing or heart-beating motion or frames of an endoscopic video show smooth non-linear variation between individual data points. Thus, in many medical applications the acquired dataset does not span the entire high dimensional image space but instead lies on a manifold of low dimension.

Since their development, the non-linear manifold learning methods [Tenenbaum et al., 2000, Roweis and Saul, 2000, Belkin and Niyogi, 2003, Coifman and Lafon, 2006] have been successfully applied to several medical applications such as classification of MR images for breathing gating [Wachinger et al., 2010] and polyps in Computed Tomography Colonography

[Suzuki et al., 2010], detection of prostate cancer in Magnetic Resonance Spectroscopy [Tiwari et al., 2008, Tiwari et al., 2009] and prostate cancer grades in MR images [Sparks and Madabhushi, 2010], content-based image retrieval of prostate histology images [Sparks and Madabhushi, 2011], segmentation [Zhang et al., 2006, Etyngier et al., 2007, Kadoury and Paragios, 2010], reconstruction [Georg et al., 2008], registration [Hamm et al., 2010, Wachinger and Navab, 2010], visualization of cardiac MR images [Souvenir and Pless, 2007], tissue characterization [Lekadir et al., 2006] and diagnosis of neural diseases [Schwarz et al., 2010]. We have also demonstrated the application of manifold learning to endoscopic datasets for clustering and classification [Atasoy et al., 2010b, Atasoy et al., 2011, Atasoy et al., 2012], which are explained in detail in Chapter 7 and Chapter 8 in this thesis, respectively. A comprehensive review of existing manifold learning techniques and their medical applications can be found in [van der Maaten et al., 2009, Pless and Souvenir, 2009, Lin and Zha, 2008] and [Mateus et al., 2012], respectively.

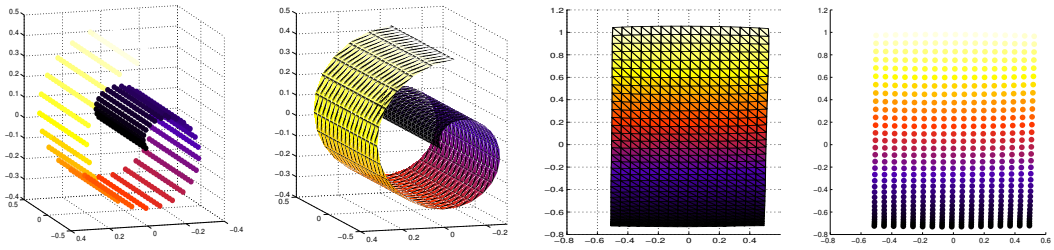
In this chapter, we will first discuss a common framework for spectral manifold learning and then review some of the most commonly used methods within this framework.

4.1 A General Recipe for Manifold Learning

A manifold can be defined as a topological space that is *locally* Euclidean. In contrast to linear vector spaces, on a manifold the Euclidean metric properties do not hold globally. Therefore, when measuring the distance between data points lying on a manifold, the global structure of the underlying manifold needs to be taken into account.

Non-linear manifold learning methods seek to find a low dimensional representation of high dimensional datasets while preserving its local structure. Embedding from the high to the low dimensional representation is computed by searching for a new coordinate system that best preserves a defined property of the underlying manifold structure. Each manifold learning method defines a different property that will be preserved in the low dimensional representation. This defined property is measured in the high dimensional space of the original observations and helps one to reveal the underlying manifold structure of this high dimensional dataset.

A clear example for manifold learning can be illustrated on the well known swiss roll dataset shown in Figure 4.1. The original input data points are represented using 3 dimensional (D) coordinates, whereas the underlying manifold is intrinsically only 2D. Embedding the data points into a lower dimensional space not only provides a dimensionality reduction but also more meaningful representation of the underlying manifold. This is because the higher dimensionality created (artificially) due to the choice of representation is now reduced to the *intrinsic* dimensionality of the manifold while preserving its actual structure. Therefore, this low dimensional space is not only more efficient due to its smaller dimensionality but it also reveals the intrinsic



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

Figure 4.1: Manifold learning demonstrated on toy swissroll example. a) Input data points in the 3D space. b) Underlying manifold structure embedded in the 3D space. c) Manifold representation in the low dimensional 2D space. d) Representation of individual data points in the low 2D space.

structure of the manifold. Thus, the low dimensional representation allows for more accurate measurements between individual data points. Interpreting a manifold as a thin metal plate, the non-linear manifold learning can be seen as expanding the plate as much as possible without creating any holes. In practice, non-linear manifold learning methods approximate the low dimensional manifold, which the data lies on, using a graph structure. The nodes of the graph represent the individual data points, whereas the edges relate to the pairwise relations between them. To this end, each data point is connected with its k-nearest neighbours *according to some similarity measure*. In fact, changing the pairwise similarities between the data points, also changes the graph structure and thus the approximated manifold. We make use of this fact for adapting the manifold structure according to the task to solve, which is demonstrated in Part III.

Let $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \in \mathcal{M} \subset \mathbb{R}^D$ be the high dimensional input data points; i.e. the observations. Manifold learning methods estimate the mapping $\boldsymbol{f} : \mathcal{M} \mapsto \mathbb{R}^d$ from the approximated manifold $\mathcal{M}$, embedded in the high dimensional space $\mathbb{R}^D$, to a low dimensional representation in $\mathbb{R}^d$ with $d << D$ for each data point $\forall \mathbf{x}_i \in \mathcal{X}$. In this new representation, the data points $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \in \mathcal{X}$ are represented with their low dimensional coordinates $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\} \in \mathcal{Y}$, where $\mathbf{y}_i = \boldsymbol{f}(\mathbf{x}_i)$. Thus, the task of manifold learning becomes equivalent to estimating the mapping function $\boldsymbol{f}$ for the input data points $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \in \mathcal{X}$.

In the literature, different manifold learning methods [Tenenbaum et al., 2000, Roweis and Saul, 2000, Belkin and Niyogi, 2003, Coifman and Lafon, 2006] propose several ways for estimating this mapping function. All existing spectral techniques; e.g. [Tenenbaum et al., 2000, Roweis and Saul, 2000, Belkin and Niyogi, 2003, Coifman and Lafon, 2006], however, can be reviewed within a common framework involving the following steps:

1. Measuring the relations between each pair of data points $\mathcal{S} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$,
2. Defining a matrix $\mathbf{H}$ *based on* these pairwise relations $\mathcal{S}(\mathbf{x}_i, \mathbf{x}_j)$, $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$,

3. Estimating the eigenvalues $\{\lambda_1, \cdots, \lambda_d\}$ and the corresponding eigenvectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ of the matrix $\mathbf{H}$.

The low dimensional representation $\mathbf{y}_i$ of a data point $\mathbf{x}_i$ is given by the i-th entry of the estimated eigenvectors[1]:

$$\boldsymbol{f}(\mathbf{x}_i) = \mathbf{y}_i = [\mathbf{v}_1(i), \cdots, \mathbf{v}_d(i)]^\top \ . \tag{4.1}$$

In order to discover the corresponding objective function, optimized by the low dimensional representation, let us first introduce the following theorem:

Theorem 4.1 (Rayleigh-Ritz theorem): *If* $\mathbf{H}$ *is an* $(n \times n)$ *Hermitian matrix* [2] *with the eigenvalues* $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$*, then for any* $(n \times 1)$ *column vector* $\mathbf{y}$ *it holds:*

$$\lambda_1 \mathbf{y}^\top \mathbf{y} \leq \mathbf{y}^\top \mathbf{H} \mathbf{y} \leq \lambda_n \mathbf{y}^\top \mathbf{y} \ . \tag{4.2}$$

For the eigenvectors $\mathbf{y}_1$ and $\mathbf{y}_n$ of the Hermitian matrix $\mathbf{H}$ corresponding to the smallest λ_1 and largest eigenvalues λ_n, it holds :

$$\begin{aligned} \mathbf{y}_1^\top \mathbf{H} \mathbf{y}_1 &= \mathbf{y}_1^\top \lambda_1 \mathbf{y}_1 = \lambda_1 \mathbf{y}_1^\top \mathbf{y}_1 \ , \\ \mathbf{y}_n^\top \mathbf{H} \mathbf{y}_n &= \mathbf{y}_n^\top \lambda_n \mathbf{y}_n = \lambda_n \mathbf{y}_n^\top \mathbf{y}_n \ . \end{aligned} \tag{4.3}$$

Expending on Equation 4.3 we can compute the following extrema values:

$$\begin{aligned} \lambda_{min} = \lambda_1 &= \min_{\mathbf{y} \neq 0} \frac{\mathbf{y}^\top \mathbf{H} \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} \ , \\ \lambda_{max} = \lambda_n &= \max_{\mathbf{y} \neq 0} \frac{\mathbf{y}^\top \mathbf{H} \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} \ . \end{aligned} \tag{4.4}$$

Thus, the eigenvectors $\mathbf{y}_1$ and $\mathbf{y}_n$ of $\mathbf{H}$ corresponding to the smallest and largest eigenvalues λ_1 and λ_n provide solutions for the following extrema problem:

$$\operatorname*{argmin}_{\mathbf{y} \neq 0} \frac{\mathbf{y}^\top \mathbf{H} \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} \ , \quad \text{and} \quad \operatorname*{argmax}_{\mathbf{y} \neq 0} \frac{\mathbf{y}^\top \mathbf{H} \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} \ , \tag{4.5}$$

[1] Depending on the defined property to preserve either the eigenvectors corresponding to the smallest d or the largest d eigenvalues are used. This is determined by the optimized objective function as introduced below.

where the extrema values correspond to the eigenvalues $\lambda_{min} = \lambda_1$ and $\lambda_{max} = \lambda_n$, respectively. For proof, we refer to [Roger and Johnson, 1990]. Therefore, the low dimensional representation defined by Equation 4.1 optimizes an objective function as defined in Equation 4.5.

Individual manifold learning techniques differ in the way they define the matrix $\mathbf{H}$ based on the pairwise relations, and thus in the preserved property in the low dimensional representations $[\mathbf{y}_1, \cdots, \mathbf{y}_d]^\top$. An insight into why a certain matrix $\mathbf{H}$ leads to the preservation of a particular property is in detail discussed in Chapter 5. As next, we introduce some of the commonly used manifold learning techniques and discuss the definition of the matrix $\mathbf{H}$ for each method.

4.2 Principal Component Analysis (PCA):

Principle component analysis (PCA) can be seen as a *linear* manifold learning technique within the above introduced general framework. Given the high dimensional data points $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \in \mathcal{X}$, PCA reduces the dimensionality of the data by finding linear combinations, called principal components, that *best preserve the variance of the dataset* $\mathcal{X}$. To this end, first a function basis $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ is computed as the eigenvectors of the covariance matrix $\mathbf{\Sigma}$ of the dataset $\mathcal{X}$ which correspond to the largest eigenvalues. These eigenvectors express the direction of maximum variance of the dataset $\mathcal{X}$. Thus, using these eigenvectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ as the new coordinate system, the data set $\mathcal{X}$ can be projected into the d dimensional space, that best preserves the variance of $\mathcal{X}$.

Let the data matrix $\mathbf{X}$ be defined as:

$$\mathbf{X} = [\overline{\mathbf{x}}_1, \cdots, \overline{\mathbf{x}}_N]^\top \ , \tag{4.6}$$

where $\overline{\mathbf{x}}_i$ is the normalized data point with zero mean $\overline{\mathbf{x}}_i = \mathbf{x}_i - \overline{\mathbf{x}}$ and $\overline{\mathbf{x}}$ denotes the mean of the data points $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \in \mathcal{X}$.

The matrix $\mathbf{H}$ in the above introduced framework is defined as the covariance matrix of the dataset $\mathcal{X}$:

$$\mathbf{H} = \mathbf{\Sigma} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top \ , \tag{4.7}$$

and provides a measure of *linear* relationship between two data points, i.e. it states how much linearly dependent the data points are on each other. Thus, its eigenvectors respect these linear relations within the dataset.

The d eigenvectors $\{\mathbf{v}_n, \cdots, \mathbf{v}_{n-d}\}$ of $\mathbf{\Sigma}$ with the largest eigenvalues $\lambda_n, \cdots, \lambda_{n-d}$, solve the following objective function:

$$\underset{\mathbf{v} \neq 0}{\operatorname{argmax}} \frac{\mathbf{v}^\top \mathbf{\Sigma} \mathbf{v}}{\mathbf{v}^\top \mathbf{v}} \ , \tag{4.8}$$

and thus provide a set of orthogonal functions which maximize the variance of the data set $\mathcal{X}$. Finally the low dimensional representation of a data points $\mathbf{x}_i$ is computed by

projecting the $\mathbf{x}_i$ into this coordinate system:

$$\mathbf{y}_i^{\text{PCA}} = \mathbf{P}\mathbf{x}_i \quad \text{with} \quad \mathbf{P} = [\mathbf{v}_n, \cdots, \mathbf{v}_{n-d}]^\top \ . \tag{4.9}$$

Note that in PCA, the eigenvectors of the covariance matrix give the function basis onto which the high dimensional data points are projected. In non-linear manifold learning methods, on the other hand, the eigenvectors provide directly the low dimensional representations. This difference is due to the definition of the optimized problem and the preserved property. The eigenvectors of the covariance matrix preserve the directions of the maximum variance and therefore are used as the coordinate system onto which the data is projected. Non-linear manifold learning methods define the property of the manifold which will be preserved in the low dimensional representation. Therefore, the eigenvectors of the defined matrix $\mathbf{H}$ provide the low dimensional representations $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ which preserve that property. The actual mapping function f from the high to the low dimensional representation is not explicitly estimated in these non-linear techniques.

4.3 ISOMAP

ISOMAP is one of the first manifold learning methods proposed by Tenenbaum *et al.* in [Tenenbaum et al., 2000]. The idea behind ISOMAP is to preserve the geodesic distances between data points after mapping into the low dimensional representation. Thus, the new coordinates are chosen to minimize the difference between the Euclidean distances in the low dimensional space and the geodesic distances in the high dimensional input space.

First, geodesic distances are measured on the manifold $\mathcal{M}$ in the high dimensional observation space. To this end, each data point is connected to its k nearest neighbours (NN) according to the chosen similarity measure $\mathcal{S}$, as introduced in Section 4.1 and a weighted graph structure is constructed. In their original paper, Tenenbaum *et al.* use the Euclidean distance as a dis-similarity measure while choosing the k-NN. Then, geodesic distance g_{ij} between each pair of data points $(\mathbf{x}_i, \mathbf{x}_j)$, $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$ is computed using the shortest path distance on the graph and encoded into the matrix $\mathbf{G} = [g_{ij}]$. The matrix $\mathbf{H}$ is defined after applying a normalization operator $\tau(\cdot)$ to the geodesic distance matrix $\mathbf{G}$ in order to convert the distances to inner products[3]:

$$\mathbf{H} = \tau(\mathbf{G}) \ . \tag{4.10}$$

The d eigenvectors of $\tau(\mathbf{G})$ with the largest eigenvalues optimize the following objective function:

$$\underset{\mathbf{v} \neq 0}{\text{argmax}} \ \frac{\mathbf{v}^\top \tau(\mathbf{G}) \mathbf{v}}{\mathbf{v}^\top \mathbf{v}} \ , \tag{4.11}$$

[3] $\tau(\mathbf{G})$ converts distance of the form $||\mathbf{v}_i - \mathbf{v}_j||^2$ into a centred dot product $\langle (\mathbf{v}_i - \overline{\mathbf{v}_i}), (\mathbf{v}_j - \overline{\mathbf{v}_j}) \rangle$, where $\overline{\mathbf{v}_i}$ denotes the average value of $\mathbf{v}_i$ [Tenenbaum et al., 2000, Bengio et al., 2004b].

and the low dimensional coordinates are given by the i-th entry of the eigenvectors:

$$\mathbf{y}_i^{\text{ISOMAP}} = \left[\sqrt{\lambda_n}\mathbf{v}_n(i), \cdots, \sqrt{\lambda_{n-d}}\mathbf{v}_{n-d}(i)\right]^\top . \tag{4.12}$$

4.4 Locally Linear Embedding (LLE)

Locally Linear Embedding (LLE) is the other pioneer work for manifold learning published in 2000 [Roweis and Saul, 2000]. LLE solves the non-linear dimensionality reduction problem by finding new coordinates which *best preserve the local reconstruction weights* of each data point. The algorithm relies on the assumption that the manifold structure is locally Euclidean and therefore each data point can be reconstructed as a linear combination of its local neighbours. The low dimensional coordinates, which best preserve these reconstruction weights, are found by minimizing the following cost function:

$$\underset{\mathbf{y}}{\text{argmin}} \sum_{\mathbf{y}_i} \left|\left| \mathbf{y}_i - \sum_{\mathbf{y}_j \in \mathcal{N}(\mathbf{y}_i)} w_{ij}\mathbf{y}_j \right|\right|^2 , \tag{4.13}$$

where w_{ij} denotes the reconstruction weight of $\mathbf{x}_i$ from $\mathbf{x}_j$ and $\mathcal{N}(\mathbf{y}_i)$ denotes the local neighbourhood of $\mathbf{y}_i$. Defining the matrix $\mathbf{H}$ as:

$$\mathbf{H} = (\mathbf{I} - \mathbf{W})^\top(\mathbf{I} - \mathbf{W}) , \quad \text{with} \quad \mathbf{W} = [w_{ij}] \tag{4.14}$$

the eigenvectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ corresponding to the smallest non-zero eigenvalues $\{\lambda_1, \cdots, \lambda_d\}$ of $\mathbf{H}$ solve the objective function in Equation 4.13 or equivalently:

$$\underset{\mathbf{v}\neq 0}{\text{argmin}} \frac{\mathbf{v}^\top(\mathbf{I} - \mathbf{W})^\top(\mathbf{I} - \mathbf{W})\mathbf{v}}{\mathbf{v}^\top\mathbf{v}} . \tag{4.15}$$

The low dimensional coordinates of each point $\mathbf{x}_i$ corresponds to:

$$\mathbf{y}_i^{\text{LLE}} = [\mathbf{v}_1(i), \cdots, \mathbf{v}_d(i)]^\top . \tag{4.16}$$

4.5 Laplacian Eigenmaps (LE)

Laplacian Eigenmaps (LE) method, which we use to create the Endoscopic Video Manifolds (EVMs) (as explained in Section 6), is introduced by Belkin and Noyagi in [Belkin and Niyogi, 2003]. The idea behind the LE method is to find a low dimensional representation *that best preserves the local structure of the manifold.* Drawing on the correspondence between the graph Laplacian $\mathbf{L}$ and the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ on the manifold $\mathcal{M}$, the matrix $\mathbf{H}$ is chosen as the graph Laplacian:

$$\mathbf{H} = \mathbf{L} , \tag{4.17}$$

and the eigenvectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ corresponding to d smallest non-zero eigenvalues $\{\lambda_1, \cdots, \lambda_d\}$ of $\mathbf{L}$ are used to compute the low dimensional coordinates:

$$\mathbf{y}_i^{\text{LE}} = [\mathbf{v}_1(i), \cdots, \mathbf{v}_d(i)]^\top \ . \tag{4.18}$$

Thus, the Laplacian eigenmaps method estimates the low dimensional coordinates by optimizing the following objective function:

$$\underset{\mathbf{y} \neq 0}{\operatorname{argmin}} \frac{\mathbf{y}^\top \mathbf{L} \mathbf{y}}{\mathbf{y}^\top \mathbf{y}} \ . \tag{4.19}$$

An interpretation for the Laplacian eigenmaps method and the property it preserves is discussed in detail in Section 5.5.

4.6 Conclusions

This chapter reviewed some of the most commonly used manifold learning methods [Tenenbaum et al., 2000, Roweis and Saul, 2000, Belkin and Niyogi, 2003] within a common mathematical framework. Further discussion on the objective functions defined by different methods and the corresponding properties that are preserved in the low dimensional representation was presented. It was pointed out that all the reviewed methods find a global solution to their optimization problem by relying on the spectral theorem.

In Chapter 5 we establish the connections between the discrete and the continuous domains and discuss an example application of the framework introduced in this Chapter in light of well studied physical phenomenon of stationary waves.

Chapter 5

Theoretical Insights

'Nature laughs at the difficulties of integration.'

Pierre-Simon Laplace

Spectral manifold learning techniques, which are the focus of this thesis, are based on solving an eigenvalue problem. In this chapter, we introduce the necessary background for the spectral theorem and eigenvalue problems. Firstly, we derive the relations between matrices, linear operators, kernel functions and Hilbert spaces. Based on these relations, we reformulate the problem solved in spectral manifold learning methods. Mathematical equivalence between the particular manifold learning method used in this thesis, the Laplacian eigenmaps, and the well-studied physical phenonemon of stationary waves is pointed out in order to present an intuitive interpretation.

5.1 From Discrete to Continuous Domain

In Chapter 4, we have showed that most spectral manifold learning methods optimize an objective function of the form:

$$\underset{\mathbf{v}\neq 0}{\operatorname{argmin}} \frac{\mathbf{v}^\top \mathbf{H}\mathbf{v}}{\mathbf{v}^\top \mathbf{v}} \ , \quad \text{or} \quad \underset{\mathbf{v}\neq 0}{\operatorname{argmax}} \frac{\mathbf{v}^\top \mathbf{H}\mathbf{v}}{\mathbf{v}^\top \mathbf{v}} \ , \tag{5.1}$$

where the matrix $\mathbf{H}$ is defined differently by individual manifold learning techniques. In this chapter, we discuss the intuition behind optimizing the quantity:

$$\mathbf{v}^\top \mathbf{H}\mathbf{v} \ . \tag{5.2}$$

In the rest of the Chapter 5, the optimized objective function in Equation 5.1 will be referred as:

$$\underset{\mathbf{v}\neq 0}{\operatorname{argmin}} \frac{\mathbf{v}^\top \mathbf{H}\mathbf{v}}{\mathbf{v}^\top \mathbf{v}} \ , \tag{5.3}$$

whereas depending on the chosen manifold learning method as discussed in Chapter 4, the argmin can be replaced by the argmax operator.

In order to provide an interpretation, we will draw on the equivalence of this problem in the continuous domain and the connections between operators, kernels and Hilbert spaces. Let us now move from the discrete to the continuous setting.

In application of manifold learning for data processing, the observations (such as images) are usually described as *vectors*:

$$\mathbf{f}, \mathbf{g} \in \mathbb{R}^n \ . \tag{5.4}$$

For the particular application addressed in this thesis, each endoscopic frame is represented using a $(w \times h)$-dimensional vector $\mathbf{I}_i \in \mathbb{R}^{(w \times h)}$, where w and h denote the width and height of the image. These observation vectors $\mathbf{f}, \mathbf{g}$ can be seen as discrete measurements

$$\mathbf{f}, \mathbf{g} : \{1, \cdots n\} \mapsto \mathbb{R} \quad \text{with} \quad n \in \mathbb{N} \tag{5.5}$$

taken from underlying continuous *functions*:

$$\boldsymbol{f}, \boldsymbol{g} : (\boldsymbol{\Omega} \subset \mathbb{R}) \mapsto \mathbb{R} \ , \tag{5.6}$$

defined on an open subset in the Euclidean space $\boldsymbol{\Omega} \subset \mathbb{R}$; i.e. an endoscopic image can be interpreted as $(w \times h)$ pixel intensities caused by the imaging of a continuous surface.

Let us now introduce a continuous linear *operator*

$$\boldsymbol{\mathcal{H}} : \mathscr{C}^\infty(\boldsymbol{\Omega}) \mapsto \mathscr{C}^\infty(\boldsymbol{\Omega}) \ , \tag{5.7}$$

where $\mathscr{C}^\infty(\boldsymbol{\Omega})$ denotes infinitely differentiable functions defined on $\boldsymbol{\Omega}$. In the continuous setting, the operator $\boldsymbol{\mathcal{H}}$ is the correspondence of $(n \times m)$ dimensional *matrix*:

$$\mathbf{H} : (\{1, \cdots, n\} \times \{1, \cdots, m\}) \mapsto \mathbb{R} \ , \quad \text{with} \quad n, m \in \mathbb{N} \ . \tag{5.8}$$

Let us further define the *inner product* $\langle \cdot, \cdot \rangle_{L2}$ between two integrable functions $\boldsymbol{f}, \boldsymbol{g} : \boldsymbol{\Omega} \mapsto \mathbb{R}$ as:

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{L2} := \int \boldsymbol{f}(t)\boldsymbol{g}(t)\mathrm{d}t \ , \tag{5.9}$$

where $\int$ denotes the Lebesgue integral. This, in tern, induces the (L2) norm in the continuous domain as:

$$\|\boldsymbol{f}\|_{L2} := \langle \boldsymbol{f}, \boldsymbol{f} \rangle_{L2} \ . \tag{5.10}$$

The equivalence of the inner product and the norm in the discrete setting are given by:

$$\langle \mathbf{f}, \mathbf{g} \rangle_{L2} := \sum_i \mathbf{f}_i \mathbf{g}_i = \mathbf{f}^\top \mathbf{g} \tag{5.11}$$

and

$$\|\mathbf{f}\|_{L2} := \mathbf{f}^\top \mathbf{f} \ . \tag{5.12}$$

A continuous kernel function, defined as:

$$\mathcal{K} : \boldsymbol{\Omega} \times \boldsymbol{\Omega} \mapsto \mathbb{R} \tag{5.13}$$

can also be described using a matrix representation $\mathbf{K} = [k_{ij}]$ in the discrete domain. Thereby, the entry for the i-th row and j-th column k_{ij} of matrix $\mathbf{K}$, corresponds to the kernel $\mathcal{K}(\boldsymbol{f}, \boldsymbol{g})$ evaluated on the functions $\boldsymbol{f}$ and $\boldsymbol{g}$ with i-th and j-th data points corresponding to the vectors $\mathbf{f}$ and $\mathbf{g}$, respectively.

A summary of the introduced correspondences in the discrete and continuous domains is presented in Table 5.1.

Discrete Domain		**Continuous Domain**	
Vectors:	$\mathbf{f}, \mathbf{g} : N \mapsto \mathbb{R}$ $N = \{1, \cdots n\},\ n \in \mathbb{N}$	Functions:	$\boldsymbol{f}, \boldsymbol{g} : \boldsymbol{\Omega} \mapsto \mathbb{R}$ $\boldsymbol{\Omega} \subset \mathbb{R}$
Matrix:	$\mathbf{H} : N \times M \mapsto \mathbb{R}$ $M = \{1, \cdots m\},\ m \in \mathbb{N}$	Operator:	$\boldsymbol{\mathcal{H}} : \mathscr{C}^\infty(\boldsymbol{\Omega}) \mapsto \mathscr{C}^\infty(\boldsymbol{\Omega})$
Matrix: (symmetric)	$\mathbf{K} : N \times N \mapsto \mathbb{R}$	Kernel:	$\mathcal{K} : \boldsymbol{\Omega} \times \boldsymbol{\Omega} \mapsto \mathbb{R}$
Inner Product:	$\langle \mathbf{f}, \mathbf{g} \rangle_{L2} = \mathbf{f}^\top \mathbf{g}$	Inner Product:	$\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{L2} = \int \boldsymbol{f}(t)\boldsymbol{g}(t)\mathrm{d}t$
Norm (L2):	$\|\mathbf{f}\|_{L2} = \langle \mathbf{f}, \mathbf{f} \rangle_{L2}$	Norm (L2):	$\|\boldsymbol{f}\|_{L2} = \langle \boldsymbol{f}, \boldsymbol{f} \rangle_{L2}$

Table 5.1: Summary of the introduced notations as correspondences in the discrete and the continuous domains.

5.2 From Linear Operators to Kernel Functions

In order to derive the equivalence of Equation (5.3) in the continuous domain, we will rely on the relations between linear operators, kernel functions and Hilbert spaces. According to the definitions in Section 5.1, a matrix in the discrete domain can be equivalent to a kernel function or a linear operator in the continuous domain. In this Section we will introduce the unique link between linear operators and kernel functions. Let us first define a linear operator $\boldsymbol{\mathcal{H}}$, which represents the equivalent of $\mathbf{H}$ in the continuous domain.

5.2.1 Continuous Linear Operators

An operator $\boldsymbol{\mathcal{H}}$ can be interpreted as the rule (or set of rules) to transform a given function $\boldsymbol{f}$ into another function $\boldsymbol{h}$. As a result, the operator $\boldsymbol{\mathcal{H}}$ applied to a function $\boldsymbol{f}$ results in another function $(\boldsymbol{\mathcal{H}f})(t) = \boldsymbol{h}(t)$.

Definition 5.1 *An operator is said to be* linear *if the following conditions hold:*

$$\begin{aligned} (\boldsymbol{\mathcal{H}}(\alpha \boldsymbol{f})) &= \alpha(\boldsymbol{\mathcal{H}f}) \ , \\ (\boldsymbol{\mathcal{H}}(\boldsymbol{f}+\boldsymbol{g})) &= (\boldsymbol{\mathcal{H}f}) + (\boldsymbol{\mathcal{H}g}) \end{aligned} \tag{5.14}$$

for all $\boldsymbol{f}, \boldsymbol{g} : \boldsymbol{\Omega} \mapsto \mathbb{R}$ *and* $\alpha \in \mathbb{R}$.

Definition 5.2 *An operator* $\boldsymbol{\mathcal{H}}$ *is* continuous *if and only if there exist a* $C > 0$ *with:*

$$||\boldsymbol{\mathcal{H}f}|| \leq C||\boldsymbol{f}|| \quad \forall \boldsymbol{f} : \boldsymbol{\Omega} \mapsto \mathbb{R} \ . \tag{5.15}$$

Drawing on the associated continuous operator $\boldsymbol{\mathcal{H}}$ of the matrix $\mathbf{H}$ in Equation (5.2) and using Equations (5.9) and (5.12), the correspondence of the optimized objective function in Equation (5.3) can be defined as follows in the continuous domain:

$$\boldsymbol{f}^* = \underset{\|\boldsymbol{f}\|_{L2} \neq 0}{\operatorname{argmin}} \frac{\langle \boldsymbol{f}, \boldsymbol{\mathcal{H}f} \rangle_{L2}}{\|\boldsymbol{f}\|_{L2}} \ . \tag{5.16}$$

5.2.2 Kernel Functions

A kernel $\mathcal{K}(t, s)$ is defined as a symmetric[1], real valued function of two variables $t, s \in \boldsymbol{\Omega}$. It assigns a value in $\mathbb{R}$ for each pair (t, s) of elements from a given space $\boldsymbol{\Omega}$:

$$\mathcal{K} : \boldsymbol{\Omega} \times \boldsymbol{\Omega} \mapsto \mathbb{R} \ , \quad (t, s) \mapsto \mathcal{K}(t, s) \ . \tag{5.17}$$

For several applications, including the kernel methods which has received significant attention in the machine learning community [Scholkopf, B., Tsuda, K., Vert, 2004, Lanckriet, G. R. G., Cristianini, N., Bartlett, P. L., Ghaoui, L. E., Jordan, 2004, Herbrich, 2002, Joachims, 2002], a kernel can be interpreted as a function evaluating the similarity (respectively distance) between each pair of objects from a given space. Further discussions on kernels and their use in the kernel methods is presented in Section 5.3.

[1] For symmetric functions it holds $\mathcal{K}(x, y) = \mathcal{K}(y, x)$.

Besides the interpretation of the kernels as similarity (respectively distance) functions, they also play an important role because of their association with continuous linear operators.

5.2.3 Kernels and Integral Operators

Schwartz's kernel theorem states the unique relation between kernel functions and continuous linear operators.

Theorem 5.1 (Schwartz's kernel theorem): *Every continuous linear operator* $\boldsymbol{\mathcal{H}}$ *is given by integration against a* unique *kernel* $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$.

$$(\boldsymbol{\mathcal{H}}\boldsymbol{f})(t) = \int \mathcal{K}_{\boldsymbol{\mathcal{H}}}(t,s)\boldsymbol{f}(s)ds \ . \tag{5.18}$$

This kernel function corresponds to the impulse response of the linear operator (linear system) to a delta function $\mathcal{K}_{\boldsymbol{\mathcal{H}}}(\cdot,s) =$ *"*$(\boldsymbol{\mathcal{H}}\delta_s)(\cdot)$*" in the sense of distributions [Schwartz, 1950].*

Thus, the objective function in Equation 5.16 can now also be written as:

$$\boldsymbol{f}^* = \underset{\|\boldsymbol{f}\|_{L2}\neq 0}{\operatorname{argmin}} \frac{\left\langle \boldsymbol{f}, \int \mathcal{K}_{\boldsymbol{\mathcal{H}}}(\cdot,s)\boldsymbol{f}(s)\mathrm{d}s \right\rangle_{L2}}{\|\boldsymbol{f}\|_{L2}} \ . \tag{5.19}$$

5.2.4 An Interpretation for Kernel Functions

For linear differential operators, the kernel function $\mathcal{K}_{\boldsymbol{\mathcal{H}}}(t,s)$ is also known as the Green's function or impulse response function of the operator $\boldsymbol{\mathcal{H}}$. Green's functions are used to describe several well studied physical phenomena; such as the diffusion of heat (diffusion equation) or the vibrations / standing waves (Helmholtz equation). In the heat equation, the Green's function $\mathcal{G}_{\boldsymbol{\mathcal{D}}}(t,s)$, which is equivalent to the Gaussian kernel, describes the temperature at point t that is created by locating a heat source $\delta(s)$ at point s. Equivalently, in describing the standing waves on a vibrating medium, the Green's function $\mathcal{G}_{\boldsymbol{\mathcal{W}}}(t,s)$ gives the amplitude of the vibration at point t for an impulse $\delta(s)$ applied at point s.

Thus, the relations between t and s encoded in the heat $\mathcal{G}_{\boldsymbol{\mathcal{D}}}(t,s)$ (or wave $\mathcal{G}_{\boldsymbol{\mathcal{W}}}(t,s)$) kernel can be interpreted as the effect of an impulse at the point s onto the point t in the diffusion (or the vibration) process.

5.3 From Kernel Functions to Hilbert Spaces

Kernels associated with positive definite linear operators are of particular interest due to their connections with Hilbert spaces.

Definition 5.3 *An operator is defined as positive definite if:*

$$\langle \boldsymbol{f}, (\boldsymbol{\mathcal{H}}\boldsymbol{f}) \rangle_{L2} > 0, \quad \forall \boldsymbol{f} : \boldsymbol{\Omega} \mapsto \mathbb{R} \ . \tag{5.20}$$

Before we investigate this connection, let us first introduce the associated positive definite kernels.

5.3.1 Positive Definite Kernels

The term *positive definite kernel* was first introduced by Mercer to define kernels $\mathcal{K} : \boldsymbol{\Omega} \times \boldsymbol{\Omega} \mapsto \mathbb{R}$ with the following property [Mercer, 1909, Aronszajn and MA., 1950]:

$$\sum_{i,j=1}^{n} \mathcal{K}(t_i, t_j)\overline{\epsilon_i}\epsilon_j \geq 0 \quad t_i, t_j \in \boldsymbol{\Omega} \text{ and } \epsilon_i, \epsilon_j \in \mathbb{C} \ . \tag{5.21}$$

Mercer has only considered real valued kernels and thus used only real number $\epsilon_i, \epsilon_j \in \mathbb{R}$ [Mercer, 1909]. The extension for complex valued kernels with $\epsilon_i, \epsilon_j \in \mathbb{C}$ was introduced by Aronszajn [Aronszajn and MA., 1950]. In the rest of this chapter, we will restrict ourselves to real-valued, positive definite kernels. For the simplicity of the presentation, they will be referred to as kernels, or kernel functions.

5.3.2 Hilbert Spaces

Hilbert spaces can be considered as higher dimensional generalizations of the Euclidean space.

Definition 5.4 *A linear space $\mathscr{H}$ (of objects such as vectors or functions) is called a* Hilbert space*, if*

- *there exists an inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathscr{H}}$, $\quad \forall \boldsymbol{f}, \forall \boldsymbol{g} \in \mathscr{H}$ and*
- *$\mathscr{H}$ is complete with respect to the norm $||\cdot||_{\mathscr{H}}$ induced by the inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$.*

A function space $\mathscr{H}$ is said to be complete with respect to the norm $||\cdot||_{\mathscr{H}}$ if every Cauchy sequence[2] $\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots$ converges[3] to an element in $\mathscr{H}$. The completeness can be interpreted as the equivalent of a closedness for the function space.

The definition of the inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ specifies the notion of angle and the norm $||\cdot||_{\mathscr{H}}$ specifies the notion of distance[4] between two functions in this particular Hilbert space $\mathscr{H}$. However, the existence of an inner product has further theoretical consequences.

One consequence of the existence of the inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ is that it allows for the projection of the elements in this Hilbert space onto a new function basis. An element of a Hilbert space can be uniquely specified by its coordinates with respect to a defined function basis. Let the function basis be defined as $\boldsymbol{\psi} = \{\psi\}_{i=1}^{\infty}$. In our case, $\mathscr{H} = \mathscr{C}^{\infty}(\boldsymbol{\Omega})$, which implies that $\mathscr{H}$ is separable and the basis $\boldsymbol{\psi}$ is countable. Then, each element $\boldsymbol{f} \in \mathscr{H}$ can be represented as the linear combination of the basis functions:

$$\boldsymbol{f} = \sum_{i=1}^{\infty} \alpha_i \psi_i \ , \quad \alpha_i = 0 \ \text{ almost everywhere (a.e.)} \ . \tag{5.22}$$

A function basis is said to be orthonormal if $||\psi_i|| = 1$ for all i and $\langle \psi_i, \psi_j \rangle_{L2} = 0$ for all $i \neq j$. Each element of the Hilbert space $\mathscr{H}$ can be represented by its projection onto this function basis. The representation of a function $\boldsymbol{f}$ in an orthonormal basis function is given by:

$$[\langle \boldsymbol{f}, \psi_1 \rangle_{\mathscr{H}}, \langle \boldsymbol{f}, \psi_2 \rangle_{\mathscr{H}}, \cdots, \langle \boldsymbol{f}, \psi_{\infty} \rangle_{\mathscr{H}}]^{\top} , \quad \alpha_i = 0 \ \text{ a.e.} \ , \tag{5.23}$$

where α_i in Equation (5.22)equals to $\langle \boldsymbol{f}, \psi_i \rangle_{\mathscr{H}}$.

Within the spectral manifold learning framework, the high dimensional data points can be considered as elements of a Hilbert space. Thus, by defining a particular inner product formulation, manifold learning methods determine the equivalent Hilbert space. Using the eigenfunctions of the corresponding operator and equivalently of the corresponding kernel, the data points are represented in the canonical coordinates of this Hilbert space. In Section 5.4.1, we discuss the relation between the eigenfunctions of a reproducing kernel or the corresponding linear operator and the preserved quantities by several manifold learning methods.

[2] A sequence $\{\boldsymbol{f}\}_{i=1}^{\infty} \subset \mathscr{H}$ is a *Cauchy sequence* if and only if:

$$\forall \epsilon > 0, \ \exists z \in \mathbb{N} : ||\boldsymbol{f}_n - \boldsymbol{f}_m|| < \epsilon, \ \forall n, \forall m \geq z \ .$$

[3] A sequence $\{\boldsymbol{f}\}_{i=1}^{\infty} \subset \mathscr{H}$ *converges* to $\overline{\boldsymbol{f}}$ if and only if:

$$\forall \epsilon > 0, \ \exists z \in \mathbb{N} : ||\boldsymbol{f}_n - \overline{\boldsymbol{f}}|| < \epsilon, \ \forall n \geq z \ .$$

[4] Within the manifold learning framework, particularly while creating the endoscopic video manifolds, the term "distance" is not necessarily defined as the norm of the difference $||\mathbf{f} - \mathbf{g}||$ between two elements $\mathbf{f}$ and $\mathbf{g}$ but can be replaced by other metric measures. This, in fact, allows us to adapt the structure of the manifold according to the application and is discussed in detail in Chapter 7.

5.3.3 Reproducing Kernel Hilbert Spaces

Let us begin by introducing the term "reproducing kernel".

Definition 5.5 (Reproducing Kernel): *For a class of integrable functions* $\mathscr{G} = \{\boldsymbol{g} : \boldsymbol{\Omega} \mapsto \mathbb{R}\}$, *the kernel function* $\mathcal{K}(t, s)$ *is called* reproducing kernel (RK) *of* $\mathscr{G}$ *if [Aronszajn and MA., 1950]:*

$$\begin{aligned} &\forall t \in \Omega && \mathcal{K}(t, \cdot) \in \mathscr{G} \\ &\forall t \in \Omega \text{ and } \forall \boldsymbol{g} \in \mathscr{G} && \boldsymbol{g}(t) = \int \mathcal{K}(t, s)\boldsymbol{g}(s)ds \ . \end{aligned} \tag{5.24}$$

Equation 5.24 states the reproducing property, as the function $\boldsymbol{g}$ at the point t is reproduced from its values $\boldsymbol{g}(s)$ using the inner product with the kernel $\mathcal{K}$. Note that the term

$$\int \mathcal{K}(\cdot, s)\boldsymbol{g}(s)\mathrm{d}s$$

is equivalent to the inner product

$$\langle \mathcal{K}(\cdot, s), \boldsymbol{g}(s) \rangle_{L2}$$

with respect to the variable s.

Moore has showed that to each positive definite kernel corresponds a class of functions for which the kernel possess the *reproducing* property [Moore, 1916]. In other words, there exists a well determined class $\mathscr{H} = \{\boldsymbol{f} : \boldsymbol{\Omega} \mapsto \mathbb{R}\}$ of functions $\boldsymbol{f}$, in respect to which the kernel $\mathcal{K}$ has the "reproducing property" [Moore, 1916, Aronszajn and MA., 1950]. In [Aronszajn and MA., 1950], Aronszajn, has presented the following theorem (Theorem 5.2) stating the link in both directions first time; i.e. the existence of a unique Hilbert Space ($\mathscr{H}$) for every positive definite kernel ($\mathcal{K}$) and also the existence of a unique kernel for every Hilbert space.

Theorem 5.2 (Moore-Aronszajn theorem): *For each symmetric, positive definite kernel* $\mathcal{K} : \boldsymbol{\Omega} \times \boldsymbol{\Omega} \mapsto \mathbb{R}$, *there exists a* unique *class* $\mathscr{H}$ *of integrable functions* $\boldsymbol{f} : \boldsymbol{\Omega} \mapsto \mathbb{R}$, *in respect to which the kernel* $\mathcal{K}$ *has the "reproducing property" (Equation (5.24)) and vice versa [Moore, 1916, Aronszajn and MA., 1950]. The determined class of functions forms a Hilbert space* $\mathscr{H}$ *with a scalar product* $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ *[Moore, 1916]. This Hilbert space of functions is called the* Reproducing Kernel Hilbert Space (RKHS).

In terms of practical applications, kernels also play an important role for solving partial differential equations. This ties together with the connection of kernels and Green's functions (fundamental solutions of partial differential equations) as introduced in Section 5.2. For a partial differential equation (PDE), the reproducing kernel of the class of solutions is the difference of the associated Neumann's function (accounting for the particular boundary conditions) and Green's function (related to the fundamental solutions of the particular PDE) [Bergman and Schiffer, 1951, Aronszajn and MA., 1950]. The importance of the reproducing kernels and the theory behind them is nicely summarized by Larkin: "The reproducing kernel function often plays a key role in bridging the gulf separating the abstract formalism of functional analysis from the computational applications" [Larkin, 1983].

In order to see the role of RKs in non-linear manifold learning let us now introduce the inner product according to a RK $\langle \cdot, \cdot \rangle_{\mathcal{K}_{\boldsymbol{\mathcal{H}}}}$ as:

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{K}_{\boldsymbol{\mathcal{H}}}} := \iint \boldsymbol{f}(t) \mathcal{K}_{\boldsymbol{\mathcal{H}}}(t, s) \boldsymbol{g}(s) \mathrm{d}s\mathrm{d}t \tag{5.25}$$

with the inner product of the corresponding RKHS being defined as:

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathscr{H}} := \langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{K}_{\boldsymbol{\mathcal{H}}}} \ . \tag{5.26}$$

Recalling the unique relation between a linear operator $\boldsymbol{\mathcal{H}}$ and a kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ according to Schwartz's kernel theorem (Theorem 5.1), Equation (5.25) leads to

$$\begin{aligned} \langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{K}_{\boldsymbol{\mathcal{H}}}} &= \iint \boldsymbol{f}(t) \mathcal{K}_{\boldsymbol{\mathcal{H}}}(t, s) \boldsymbol{g}(s) \mathrm{d}s\mathrm{d}t \\ &= \int \boldsymbol{f}(t) (\boldsymbol{\mathcal{H}}\boldsymbol{g})(t) \mathrm{d}t \\ &= \langle \boldsymbol{f}, \boldsymbol{\mathcal{H}}\boldsymbol{g} \rangle_{L2} \ . \end{aligned} \tag{5.27}$$

Thus, in the discrete domain, Equation (5.25) corresponds to the following definition :

$$\langle \mathbf{f}, \mathbf{g} \rangle_{\mathbf{H}} = \mathbf{f}^\top \mathbf{H} \mathbf{g} \ , \tag{5.28}$$

where the matrix $\mathbf{H}$ is the analogous of the linear operator $\boldsymbol{\mathcal{H}}$ in the discrete setting.

Therefore, the Equation (5.3) which is optimized by the spectral manifold learning techniques corresponds to following norm in the continuous domain:

$$\begin{aligned} \boldsymbol{f}^* &= \underset{||\boldsymbol{f}|| \neq 0}{\operatorname{argmin}} \frac{\langle \boldsymbol{f}, \boldsymbol{\mathcal{H}}\boldsymbol{f} \rangle_{L2}}{\langle \boldsymbol{f}, \boldsymbol{f} \rangle_{L2}} \\ &= \underset{||\boldsymbol{f}|| \neq 0}{\operatorname{argmin}} \frac{\langle \boldsymbol{f}, \boldsymbol{f} \rangle_{\mathcal{K}_{\boldsymbol{\mathcal{H}}}}}{\langle \boldsymbol{f}, \boldsymbol{f} \rangle_{L2}} \\ &= \underset{||\boldsymbol{f}|| \neq 0}{\operatorname{argmin}} \frac{||\boldsymbol{f}||^2_{\mathscr{H}}}{||\boldsymbol{f}||^2_{L2}} \ , \end{aligned} \tag{5.29}$$

where the norm $||\cdot||_{\mathcal{K}_{\boldsymbol{\mathcal{H}}}} = ||\cdot||_{\mathscr{H}}$ is defined in the corresponding Hilbert space $\mathscr{H}$ of the kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ and is associated with the corresponding linear operator $\boldsymbol{\mathcal{H}}$.

Therefore, by defining a matrix $\mathbf{H}$ in the discrete domain, one equivalently determines the corresponding linear operator $\boldsymbol{\mathcal{H}}$ applied to the data points lying on the manifold, which in turn determines the corresponding kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ and hence the associated Hilbert space $\mathscr{H}$ in which the inner product is evaluated in the continuous setting. Since the denominator in Equation (5.29) does not depend on the choice of $\boldsymbol{\mathcal{H}}$ (and accordingly of $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ and $\mathscr{H}$), this can be considered as a normalisation to remove the scaling factor. Thus, by optimizing the Equation (5.3) in discrete and Equation (5.29) in continuous domains, we seek for functions $\boldsymbol{f}$ with minimal length $||\boldsymbol{f}||_{\mathscr{H}}$ in the Hilbert space $\mathscr{H}$ (after normalisation by $\|\cdot\|_{L2}$).

In conclusion, the choice of the matrix $\mathbf{H}$ within the manifold learning framework is equivalent to defining the kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ (and accordingly the operator $\boldsymbol{\mathcal{H}}$) expressing the relations between data points which are to be preserved in the new low dimensional representation. This, in turn, is equivalent to defining the Hilbert space $\mathscr{H}$ according to which we take the norm.

The next section studies the relation between the Hilbert space $\mathscr{H}$ associated to the kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ and the eigenfunctions $\{\psi_i\}_{i=0}^{\infty}$ of the linear operator $\boldsymbol{\mathcal{H}}$.

5.4 From Eigenfunctions to Feature Spaces

So far we have established the relations between linear operators, positive definite kernels and RKHS. This provides an interpretation of the objective function in Equation (5.3) used in the manifold learning framework as discussed above.

According to the Rayleigh-Ritz theorem (Theorem 4.1), the eigenvectors of the matrix $\mathbf{H}$ with smallest eigenvalues give global optimum of the Equation (5.3). Let us now seek an interpretation for the eigenvectors of the matrix $\mathbf{H}$ and eigenfunctions of the operator $\boldsymbol{\mathcal{H}}$.

5.4.1 Eigenfunctions of an Operator

The eigenvectors of a matrix $\mathbf{H} \in \mathbb{R}^{n,\times m}$, $n, m \in \mathbb{N}$ are defined as vectors that satisfy:

$$\mathbf{H}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad \mathbf{v}_i \neq 0, \quad \lambda_i \in \mathbb{C}\ . \tag{5.30}$$

Equivalently, in the continuous domain, the eigenfunctions of an operator $\boldsymbol{\mathcal{H}}$ satisfy:

$$\boldsymbol{\mathcal{H}}\boldsymbol{\psi}_i = \lambda_i \boldsymbol{\psi}_i, \quad \boldsymbol{\psi}_i \neq 0, \quad \lambda_i \in \mathbb{C}\ . \tag{5.31}$$

Considering an operator as the rules that transform a given function into another function [Levine, 1983], the eigenfunction correspond to functions that transform to themselves (more precisely to the scaled version of themselves). In a physical system

where the operator $\boldsymbol{\mathcal{H}}$ describes the evolution of the system over time, the eigenfunctions $\boldsymbol{\psi}_i$ of the operator correspond to the steady states of the system. For instance, vibrations of a medium such as a vibrating string or a vibrating membrane are described by the Laplace operator applied to that particular domain with certain boundary conditions. The eigenfunctions $\boldsymbol{\psi}_i$ of the Laplace operator on that particular domain satisfying the boundary conditions give the steady states of the vibrating system at different frequencies λ_i. This example is discussed in Section 5.5.4 in further detail.

Let us turn back to the question why the eigenfunctions $\boldsymbol{\psi}_i$ of an operator $\boldsymbol{\mathcal{H}}$ give us the optimum representation for the data, if in this new representation a certain property is to be preserved.

The spectral manifold learning techniques reviewed in Chapter 4 include the quantity to be preserved into the matrix $\mathbf{H}$; e.g. for two data points $\mathbf{x}_i$ and $\mathbf{x}_j$, the corresponding entry of the matrix $\mathbf{H} = [h_{ij}]$ denotes the covariance, geodesic distances and local reconstruction weights in PCA, ISOMAP and LLE, respectively. The matrix $\mathbf{H}$ can be seen as the operator $\boldsymbol{\mathcal{H}}$ applied to the discrete set of data points. Since the eigenfunctions (respectively eigenvectors) follow the transformation of the linear operator $\boldsymbol{\mathcal{H}}$ (respectively the matrix $\mathbf{H}$), they respect the relations expressed in the operator $\boldsymbol{\mathcal{H}}$ or the corresponding kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ (respectively the matrix $\mathbf{H}$).

Besides this direct interpretation, the eigenfunctions also play an important role in the formation of a new feature space. Next Section bridges the concepts of linear operators and positive definite kernels with feature spaces.

5.4.2 Feature Spaces

Let us introduce a function $\boldsymbol{\Phi} : \boldsymbol{\Omega} \mapsto \mathscr{F}$ mapping the input space $\boldsymbol{\Omega}$ to the feature space $\mathscr{F} = \{\boldsymbol{f} : \boldsymbol{\Omega} \mapsto \mathbb{R}\}$. For every positive definite kernel, there exists a mapping function $\boldsymbol{\Phi}$ and a feature space $\mathscr{F}$, where the kernel can be expressed in the inner product form:

$$\mathcal{K}(t,s) = \langle t,s \rangle_{\mathscr{F}} = \langle \boldsymbol{\Phi}(t), \boldsymbol{\Phi}(s) \rangle \ . \tag{5.32}$$

To see this relation we draw on the eigenfunctions of the positive definite kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ and the associated linear operator $\boldsymbol{\mathcal{H}}$. Let us first recall the Mercer-Hilbert-Schmidt theorem.

Theorem 5.3 *(**Mercer-Hilbert-Schmidt theorem**) If $\mathcal{K}_{\boldsymbol{\mathcal{H}}}(\mathrm{t},\mathrm{s})$ is a positive definite kernel, then there exists an infinite sequence of eigenvalues $\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$ of $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ and the corresponding eigenfunctions $[\boldsymbol{\phi}_i]_{i=0}^{\infty}$ such that*

$$\int \mathcal{K}_{\boldsymbol{\mathcal{H}}}(t,s)\boldsymbol{\phi}_i(t)\boldsymbol{\phi}_i(s)\,dt\,ds \geq 0 \ . \tag{5.33}$$

The eigenfunctions $\boldsymbol{\phi}_i$ form an orthonormal basis and the kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ can be expressed as:

$$\mathcal{K}_{\boldsymbol{\mathcal{H}}}(t,s) = \sum_{i=0}^{\infty} \lambda_i \boldsymbol{\phi}_i(t)\boldsymbol{\phi}_i(s) \ . \tag{5.34}$$

The eigenfunctions of a kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ are defined as the solutions $\{\boldsymbol{\phi}_i\}$ of the following equation:

$$\int \mathcal{K}_{\boldsymbol{\mathcal{H}}}(t,s)\boldsymbol{\phi}_i(s)\mathrm{d}s = \lambda_i \boldsymbol{\phi}_i(t) \ , \quad \forall t \ , \tag{5.35}$$

or equivalently of:

$$\langle \mathcal{K}_{\boldsymbol{\mathcal{H}}}(t,\cdot), \boldsymbol{\phi}_i \rangle = \lambda_i \boldsymbol{\phi}_i(t) \ , \quad \forall t \ . \tag{5.36}$$

Thus the eigenfunctions $[\boldsymbol{\phi}]_{i=1}^{\infty}$ of the kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ are equivalent to the eigenfunctions $[\boldsymbol{\psi}]_{i=1}^{\infty}$ of the corresponding integral operator $\boldsymbol{\mathcal{H}}$.

Let the mapping function $\boldsymbol{\Phi}$ from the input space $\boldsymbol{\Omega}$ to the feature space $\mathscr{F}$ be defined as:

$$\boldsymbol{\Phi}(t) := \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{\phi}_1(t) \\ \sqrt{\lambda_2}\boldsymbol{\phi}_2(t) \\ \vdots \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{\psi}_1(t) \\ \sqrt{\lambda_2}\boldsymbol{\psi}_2(t) \\ \vdots \end{bmatrix} \ . \tag{5.37}$$

Drawing on Equation (5.34) and Equation (5.37), the kernel $\mathcal{K}_{\boldsymbol{\mathcal{H}}}$ can be expressed as the inner product formulation:

$$\mathcal{K}_{\boldsymbol{\mathcal{H}}}(t,s) = \langle \boldsymbol{\Phi}(t), \boldsymbol{\Phi}(s) \rangle \ . \tag{5.38}$$

Now we have shown that every positive definite kernel (reproducing kernel) $\mathcal{K}$ is the inner product in some feature space. This, in fact, is the basis of the well known kernel-trick in the machine learning community. The basic idea behind the kernel methods is to map the (relations between the) input data points onto a new and possibly infinite dimensional feature space, in which the non-linear relations become linear. In this feature space, the tools to handle linear relations can be utilized. Furthermore, in order to use the "kernel trick" the explicit definition of the mapping $\boldsymbol{\Phi} : \boldsymbol{\Omega} \mapsto \mathscr{F}$ from the input space $\boldsymbol{\Omega}$ to a feature space $\mathscr{F}$ is not necessary. The so called kernel methods [Scholkopf, B., Tsuda, K., Vert, 2004, Lanckriet, G. R. G., Cristianini, N., Bartlett, P. L., Ghaoui, L. E., Jordan, 2004, Herbrich, 2002, Joachims, 2002] utilize only the kernel function. In [Smale et al., 2010], the authors derive a kernel function which reflects the human perception of the similarity between two images. In contrary to the above mentioned kernel methods, in this work a new mapping function from the input to the feature space is proposed. The mapping function is motivated by the neuroscience of the visual cortex. The distance reflecting human perception of image similarities is then simply formulated as the inner product in this feature space [Smale et al., 2010, Wibisono et al., 2010].

Furthermore, the correspondences introduced in this chapter also provide the direct relation between kernel methods, in particular kernel PCA [Scholkopf, B., Tsuda, K., Vert, 2004], and the spectral manifold learning methods [Tenenbaum et al., 2000, Roweis and Saul, 2000, Belkin and Niyogi, 2003]. Kernel PCA extends the linear PCA method to account for non-linear relations between data points using the kernel trick. To this end, using a kernel function the data points are intrinsically mapped into a feature space where non-linear relations become linear. The linear dimensionality reduction method, PCA, is then performed in this feature space. Due to the use of this new feature space, the linear PCA method is extended to preserve non-linear relations between data points in the low dimensional space. A detailed study on the relation of the spectral manifold learning methods to the kernel PCA can be found in [Bengio et al., 2004a, Bengio et al., 2004b].

5.5 Interpretation for Laplacian Eigenfunctions

In this Section, we review the eigenfunctions of the Laplace operator in light of the relations established in Sections 5.1, 5.2, 5.3 and 5.4. As explained in Chapter 6, for creating the EVMs we utilize the Laplacian eigenmaps method [Belkin and Niyogi, 2003] which computes the low dimensional representation of the input data points using the eigenvectors of the graph Laplacian. Supposing that the input data points are samples taken homogeneously from an underlying manifold, the graph Laplacian provides an approximation of the Laplace-Beltrami operator applied to the manifold [Belkin and Niyogi, 2007]. In order to develop an interpretation for the eigenfunctions of the Laplace and the Laplace-Beltrami operators, we will draw on their application to describe stationary (standing) waves.

5.5.1 Laplace Operator

Laplace operator Δ is defined as the combination of two differential operators; i.e. the divergence $\nabla\cdot$ of the gradient ∇ of a function:

$$\Delta := \nabla \cdot \nabla \ . \tag{5.39}$$

It is worth noting that the divergence $\nabla\cdot$ and the gradient operator ∇ are adjoint operators[5].

The Laplace operator, also referred to as the Laplacian, plays a very important role in describing several physical phenomena such as heat diffusion, electrostatistic, wave equation and quantum mechanics. Drawing on these physical examples, J. E. McDonald suggests an interpretation for the Laplace operator as the local anomalies [McDonald, 1965]. In equations describing the behaviour of these physical phenomena,

[5]Every continuous linear operator $\boldsymbol{\mathcal{A}} : \mathscr{H} \mapsto \mathscr{H}$ on a Hilbert space $\mathscr{H}$ has a unique adjoint operator $\boldsymbol{\mathcal{A}}^*$ such that $\langle \boldsymbol{\mathcal{A}} \boldsymbol{f}, \boldsymbol{g} \rangle = \langle \boldsymbol{f}, \boldsymbol{\mathcal{A}}^* \boldsymbol{g} \rangle$.

local anomalies (differences) measured by the Laplace operator applied to a distribution (such as temperature) create a potential or change of this distribution over time. This can be interpreted as describing the effect caused by the local differences in a physical system.

In this chapter, we draw on the physical example of standing waves or vibrational modes of a domain in order to gain an interpretation for the eigenfunctions of the Laplacian. First, let us introduce the extension of the Laplace operator for a manifold structure.

5.5.2 Laplace-Beltrami Operator

Laplace-Beltrami operator is the extension of the Laplacian to non-euclidean geometries such as a Riemannian manifold. In n-dimensional Euclidean space, Laplace operator is equal to the sum of all second order partial derivatives:

$$\Delta := \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i} \ , \tag{5.40}$$

with $\mathbf{x} = [x_1, \cdots, x_i, \cdots, x_n]^\top$ being the n dimensional variable. The equivalent of the Laplacian on a manifold $\mathcal{M}$ leads to the Laplace-Beltrami operator where the derivatives are computed in the corresponding tangent space of each point $\mathbf{x}$ on the manifold. For further details on the derivation of the Laplace-Beltrami operator, we refer to [Belkin and Niyogi, 2003, Belkin, 2003]. In the rest of this thesis, the Laplace and Laplace-Beltrami operators will be denoted by Δ and $\Delta_{\mathcal{M}}$, respectively.

5.5.3 Graph Laplacian

In the discrete domain, the graph Laplacian can be considered as the analogous of the Laplace-Beltrami operator on a manifold [Belkin and Niyogi, 2007]. In the spectral graph theory, the graph Laplacian, its eigenvalues and eigenfunctions have been well studied and applied in several domains [Chung, 1997].

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be described as a discrete set of vertices or nodes $\mathcal{V} = \{v_i : 1 \leq n\}$ with $n \in \mathbb{N}$ and a collection of edges; i.e. the connections between the nodes $\mathcal{E} = \{e_{i,j} | (i,j) \in \mathcal{V} \times \mathcal{V}\}$. Two vertices v_i and v_j are said to be *adjacent* if they are connected by an edge $(i,j) \in \mathcal{E}$. For a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{W}$ contains the weight of each edge, the strength of the connections is also taken into account by specifying the weight w_{ij} for each edges (i,j). The connectivity of a graph can be represented using an adjacency matrix $\mathbf{A}$, which is defined as

$$\mathbf{A}(i,j) = \begin{cases} 1 & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise,} \end{cases} \tag{5.41}$$

for an unweighted graph and as

$$\mathbf{A}(i,j) = \begin{cases} w_{ij} & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise,} \end{cases} \tag{5.42}$$

for a weighted graph.

In the context of spectral manifold learning methods, the adjacency matrix is defined based on the similarities or distances of individual data points. In fact, the similarity measure which defines the connections of a graph can be designed according to the task to perform on the manifold, such as clustering informative/uninformative frames or clustering different endoscopic segments. In Chapter 7, we introduce different similarity measures leading to very different manifold structures, each of which is more suitable for performing a different clustering task.

Given the adjacency matrix $\mathbf{A}$ of a graph $\mathcal{G}$, the *graph Laplacian* $\mathbf{L}$ is computed as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \ , \tag{5.43}$$

where $\mathbf{D}$ denotes the diagonal *degree matrix*

$$\mathbf{D}(i,i) = \sum_{j=1}^{n} \mathbf{A}(i,j) \ . \tag{5.44}$$

For an unweighted graph, where the adjacency matrix only describes the presence or absence of a connection between the vertices, the graph Laplacian is called *combinatorial graph Laplacian*. More detailed discussions on the combinatorial Laplacian are presented in Sections 6.2 and 6.7.2 in Chapter 6.

5.5.4 Eigenfunctions of the Laplace and Laplace-Beltrami Operators

In Chapter 4, we have reviewed the spectral manifold learning methods and discussed that all these methods rely on the eigenfunctions of a matrix $\mathbf{H}$ (equivalently of an operator $\boldsymbol{\mathcal{H}}$ in the continuous domain). Among other spectral manifold learning techniques, the Laplacian eigenmaps method [Belkin and Niyogi, 2003] is of particular interest due to its association with the Laplace and Laplace-Beltrami operators. Laplacian eigenmaps method computes the low dimensional coordinates of data points lying on a manifold using the eigenvectors of the graph Laplacian $\mathbf{L}$. In the continuous domain this is analogous to computing the eigenfunctions of the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$, which in turn is equivalent to solving the *Helmholtz Equation* on underlying manifold $\mathcal{M}$:

$$\Delta_{\mathcal{M}} \boldsymbol{f} = \lambda \boldsymbol{f} \ . \tag{5.45}$$

Helmholtz Equation describes the vibrational modes of a domain $\mathcal{M}$ with particular boundary conditions. Vibrational modes of a domain refer to standing (stationary)

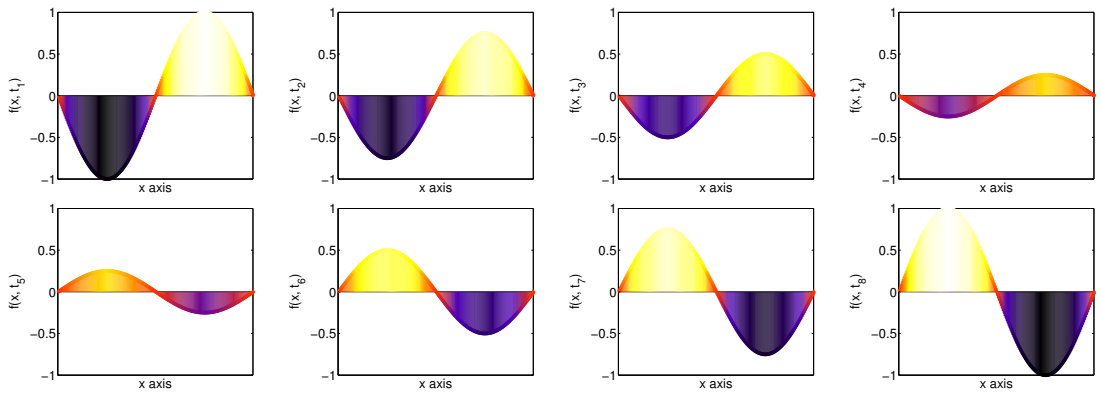


Figure 5.1: Evaluation of the first vibrational mode of a string with two fixed end points, where the x-axis represents different locations of the string and the y-axis denotes $\boldsymbol{f}(x, t_i)$, the amplitude of the vibration at point x for time t_i. Each point on the string moves up and down with a fixed maximum amplitude over time.

waves created on this medium with the corresponding boundary conditions. Several waves with the same frequency propagating through the same medium give rise to the so called standing waves. This can also be achieved by creating waves on a bounded domain, as the boundary conditions lead to the reflection of the propagating wave with the same frequency and result in standing waves. These waves are called stationary or standing as the wave profile does not propagate through the domain over time but each location moves only up and down with a fixed maximum amplitude yielding a vibrational mode as illustrated in Figure 5.1.

As next, we discuss examples for standing waves in on several domains with different dimensionality.

5.5.4.1 Vibrational Modes in 1D

If we imagine the vibrating domain to be a string fixed at both ends, such as a guitar string, solutions of the Helmholtz Equation; i.e. the eigenfunctions of the Laplace-Beltrami operator applied to this domain, yield the amplitude of the vibration at each mode. In other words, the eigenfunctions $\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots$ corresponding to the eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$ describe the amplitude of each location x in this vibrational mode. The first four vibrational modes of a vibrating string are illustrated in Figure 5.2. In the example of the vibrations of the guitar string, the corresponding eigenvalues relate to the frequency of the vibration and the sound we hear.

Solving the Helmholtz Equation on a circular domain corresponds to finding the eigenfunctions of the Laplace-Beltrami operator on this domain with cyclic boundary

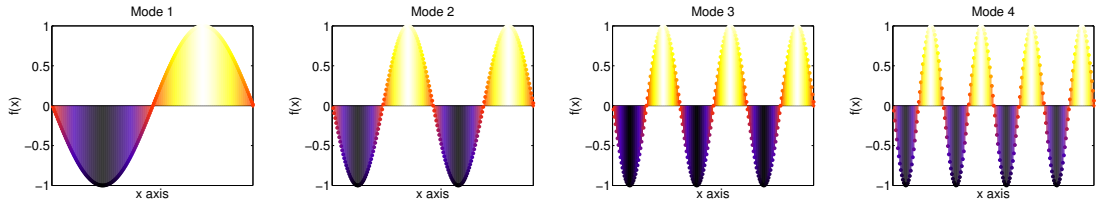


Figure 5.2: Vibrational modes of a string with two fixed end points corresponding to the 4 lowest frequencies or equivalently 4 smallest eigenvalues.

conditions [Levy, 2006]; i.e. finding $\boldsymbol{f}$ with

$$(\Delta \boldsymbol{f}) = \left(\frac{\partial^2}{\partial x_1} \boldsymbol{f} \right) = \lambda \boldsymbol{f} \ , \tag{5.46}$$

with $\boldsymbol{f} : [0, 2\pi] \mapsto \mathbb{R}$ and $\boldsymbol{f}(0) = \boldsymbol{f}(2\pi)$ which leads to $\sin(Nx)$ and $\cos(Nx)$ with $N \in \mathbb{N}$ as the eigenfunctions of the second order derivative for the given boundary conditions[6]. One can also see that the eigenfunctions of the Laplace-Beltrami operator on a circular domain construct the function basis used in the Fourier transform. This observation provides a basis to extend the Fourier transform to other domains as discussed in 5.5.4.4.

5.5.4.2 Vibrational Modes in 2D

Study of the vibrational modes of a two dimensional (2D) domain have been first published in 1787 by Ernst Chladni as "Discoveries Concerning the Theory of Music" [Ernst Florens Friedrich Chladni, 1787]. In his experiment, Chladni fixed a thin metal plate from its centre point and spread sand over the metal plate. By putting the plate into vibration, Chladni observed that the sand accumulates in certain zones of the metal plate. This is due to the standing wave patterns originated by the vibrations of the thin metal plate. In its vibration, the amplitude of certain areas of the plate are equal to 0; i.e. these zones, called the *nodal set*, are stationary in this vibrational mode and thus the sand accumulates in these regions. Although the patterns formed by the accumulation of the sand were very complex as shown in Figure 5.3, Chladni was able to calculate their exact shape using the Helmholtz Equation and the eigenfunctions of the Laplacian.

Supposing the function $\boldsymbol{f}(\mathbf{x})$ describes the amplitude of the vibration at a location $\mathbf{x}$ on the metal plate $\boldsymbol{\Omega}$, the amplitude at each point of the plate in a vibrational mode follows the Helmholtz Equation:

$$(\Delta_{\boldsymbol{\Omega}} \boldsymbol{f})(\mathbf{x}) = \lambda \boldsymbol{f}(\mathbf{x}) \ , \tag{5.47}$$

[6]Note that $\frac{\partial^2}{\partial x} \sin(\omega x) = \omega^2 \sin(\omega x)$ and $\frac{\partial^2}{\partial x} \cos(\omega x) = \omega^2 \cos(\omega x)$.

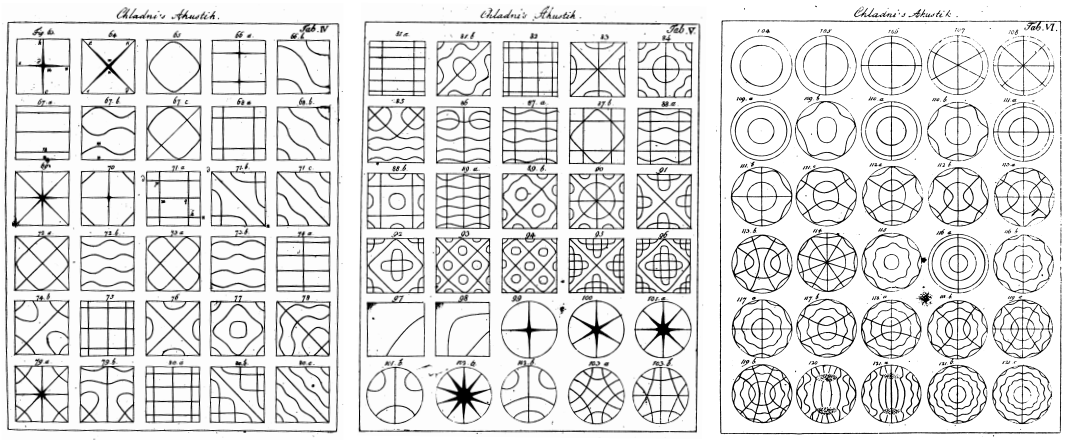


Figure 5.3: Chaldni plates. Patterns formed by the accumulation of sand at the nodal set of a vibrating metal plate. The images has been originally published in [Ernst Florens Friedrich Chladni, 1802].

where Δ_{Ω} denotes the Laplace-Beltrami operator applied to the domain Ω with the corresponding boundary conditions. Thus, the eigenfunctions of the Laplace operator (respectively Laplace-Beltrami operator on a manifold) are solutions to the Equation (5.47) and yield the amplitude of vibration for each node corresponding to the eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$. Figure 5.4 shows the first 8 vibrational modes of an elliptical domain with fixed boundaries, where the modes correspond to the lowest frequencies. Like in the 1D case, the eigenvalues relate to the sound we hear from the vibrating plate.

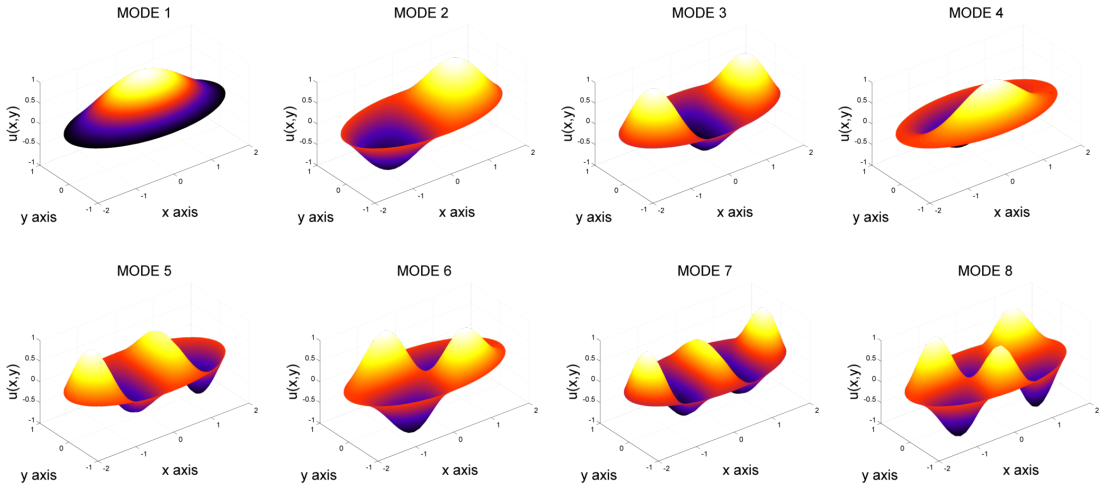


Figure 5.4: Vibrational modes of an elliptical domain with fixed boundary conditions.

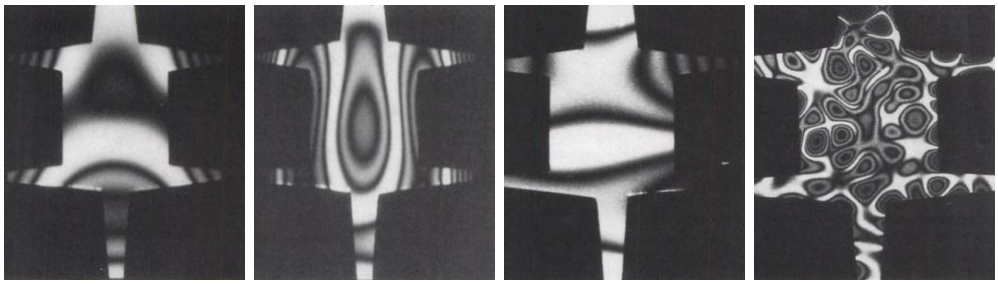


Figure 5.5: Standing wave patterns generated on a thin vibrating plate resemble animal coat patterns formed by the reaction-diffusion process, a chemical standing wave. The images have been originally published in [Murray, 1988]

This relation also led to the question *"Can we hear the shape of a drum?"* [Kac, 1966]. This question opened the discussion whether there exist a unique relation between the Laplace-Beltrami eigenfunctions (and corresponding eigenvalues) and the shape of the vibrating domain. Although it has been shown that there exists isospectral[7] drums or surfaces, spectrum of the Laplace-Beltrami operator is successfully utilized to describe shapes [Reuter et al., 2006, Rustamov, 2007] and surfaces (intensity profiles of images) [Peinecke et al., 2007]. For a more detailed study on the eigenfunctions and the nodal set of the Laplacian we refer to [Levy, 2006, Jakobson, D. and Nadirashvili, N. and Toth, 2001].

Another interesting example for patterns formed by standing waves is the formation of the animal coat patterns. In embryological morphogenesis, the spatial structure of the developing embryo is described by the principle of reaction-diffusion mechanism [Murray, 1988, Murray, 1993]. A reaction-diffusion system, which leads to a chemical standing wave, can create an infinite variety of patterns, where each pattern follows certain characteristics as observed in several animal patterns (Figure 5.5). The particular pattern is also influenced by the shape of the animal body. Murray proposed a model based on the solutions of the Helmholtz Equation to simulate animal coat patterns using the reaction-diffusion system [Murray, 1988, Murray, 1993]. Once again, although the structure of the formed patterns are very complex, the underlying mechanism can be explained by simple chemical standing waves formed by the reaction-diffusion principle.

5.5.4.3 Vibrational Modes in 3D

A well-known example for the standing wave patterns also exists for the 3D domain. The eigenfunctions of the Laplace-Beltrami operator on a sphere result in spherical harmonics [Levy, 2006, Jakobson, D. and Nadirashvili, N. and Toth, 2001], which are successfully used as a basis for shape representation. Figure 5.6 illustrates the spherical

[7] Two surfaces are called isospectral if the spectra of the Laplace-Beltrami operator on these domains are equivalent.

harmonics; i.e. the first 10 eigenfunctions of the Laplace-Beltrami operator on a sphere. The amplitude of the vibration at each vibrational mode of the sphere is indicated by the colour at this location. Drawing on the analogy between the Laplacian eigenfunctions on a circular domain, giving rise to Fourier transform, and on a sphere, the spherical harmonics can be interpreted as the three dimensional extension of the function basis used in the Fourier transform. Thus, a projection onto this function basis results in a spectral transform performed in 3D.

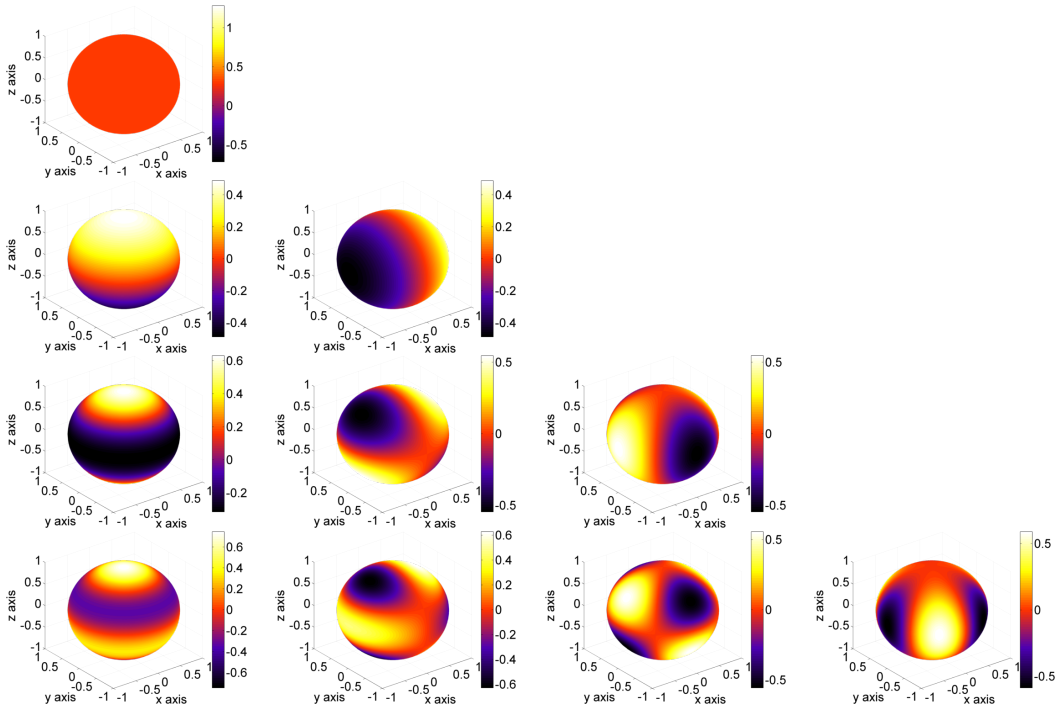


Figure 5.6: Vibrational modes of a spherical domain leading to spherical harmonic basis. The amplitude of the vibration at each location on the sphere is indicated by its colour.

5.5.4.4 Vibrating Manifolds

Drawing on the analogy of the Fourier transform, spherical harmonics and eigenfunctions of the Laplace and Laplace-Beltrami operators, it is possible to extend the spectral transform to other topologies such as manifolds. In [Vallet and Lévy, 2008], the authors propose an extension of the spectral transform for 3D meshes which approximate 3D manifold structures. To this end, eigenfunctions of the Laplace-Beltrami operator on the manifold are computed and used as the new spectral function basis, referred to as *manifold harmonics*. The transform from the Euclidean coordinates to this new spectral function basis is achieved by interpreting each dimension of the complete dataset

$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^\top$ as an n-dimensional vector and projecting each row i (corresponding to the i-th dimension all data points) separately onto the new harmonic basis.

The interpretation of the Laplace-Beltrami eigenfunctions as the vibrational modes of a manifold also provides a new insight to the Laplacian eigenmaps method [Belkin and Niyogi, 2003]. Supposing that the manifold is a thin metal plate that is curved and embedded into a higher dimensional space, its vibrational modes remain the same if the metal plate is stretched and embedded into a lower dimensional representation. Using the swiss roll example shown in Figure 4.1, one can see that the vibrational modes of the plate remain invariant if it is bent as in Figure 4.1(b) or stretched as in Figure 4.1(c). This is also in agreement with the interpretation of non-linear manifold learning as bending and stretching of a domain without creating any holes. Thus, the property respected by the Laplacian eigenmaps method is the amplitude of each point on the domain in a vibrational mode.

5.6 Conclusions

In this chapter, we provided a theoretical background for spectral manifold learning techniques. We established the relations between linear operators, kernel functions and reproducing kernel Hilbert spaces. Drawing on these relations, the objective function minimized by the spectral manifold learning methods becomes a norm in a Hilbert space. Furthermore, we derived the relation between a kernel function and a feature space, which also provides the basis for kernel methods and provides a common framework for the kernel methods and non-linear manifold learning. Finally, we discussed an interpretation for the eigenfunctions of the Laplace and Laplace-Beltrami operators in light of the derived relations. Based on the fact that these eigenfunctions are the solutions of Helmholtz Equation, we concluded that they represent the vibrational modes of a domain.

In the next chapter we present the details of our method for creating the endoscopic video manifolds.

Chapter 6

Creating Endoscopic Video Manifolds

'A man with a new idea is a crank until he succeeds.'

Mark Twain

Using the original data representation, an endoscopic video is represented as a set of frames $\mathbf{I}^{\text{RGB}} = \{\mathbf{I}_1^{\text{RGB}}, \mathbf{I}_2^{\text{RGB}}, \cdots, \mathbf{I}_n^{\text{RGB}}\} \in \mathbb{R}^{(3 \times w \times h)}$, where each frame corresponds to a 3 channel RGB (red-green-blue) colour image. This representation relies on $(3 \times w \times h)$ variables (number of pixels) to represent one endoscopic frame, where w and h denote the width and height of the frames. Even if the endoscopic frames are converted to gray-scale representation $\mathbf{I} = \{\mathbf{I}_1, \mathbf{I}_2, \cdots, \mathbf{I}_n\} \in \mathbb{R}^{(w \times h)}$, the dimensionality of an endoscopic video remains to be $(w \times h \times n)$ where n equals the number of frames. In this representation, which will be referred to as the *original image representation* in the rest of this thesis, each image can be considered as a data point in this high dimensional input space $\mathbf{I} \in \mathbb{R}^{(w \times h)}$.

If no additional constraints on the image content exist, the set of all possible images in this representation spans this high dimensional vector space $\mathbb{R}^{(w \times h)}$. Operations in such high dimensional spaces suffer from the *"curse of dimensionality"*. Similarity based retrieval, clustering and classification do not lead to meaningful results in these vector spaces due to the high dimensionality of the data points. In their influential paper titled "When is the 'nearest-neighbour' meaningful?", Beyer *et al.* demonstrate that with increasing dimensionality the distance to the nearest point approaches the distance to the farthest one [Beyer et al., 1999]. Therefore, a nearest neighbour (NN) search becomes meaningless when performed in such high dimensional spaces.

In upper GI endoscopic imaging, frames $\mathbf{I} = \{\mathbf{I}_1, \mathbf{I}_2, \cdots, \mathbf{I}_n\} \in \mathbb{R}^{(w \times h)}$ of an endoscopic video $\mathbf{I}$ do not span this high dimensional space completely but lie on (or near) a low dimensional manifold. This is due to the temporal continuity of endoscopic videos

and the large similarity between frames showing the same anatomical region. The smooth temporal change in the content of endoscopic frames results in large overlap between several consecutive frames. Furthermore, in upper GI endoscopic examinations the endoscope is first guided downwards from the mouth to the stomach and then upwards from the stomach back to the mouth. Therefore, several anatomical regions are visited multiple times during the examination. This also results in the acquisition of several visually similar frames. Thus, the frames of an endoscopic video do not span the high dimensional vector space as used in the original image representation but lie on a lower dimensional manifold embedded in this high dimensional space. The intrinsic dimensionality of the endoscopic video is much smaller than the number of degrees of freedom (DoF) in the original image representation ($w \times h$).

In this thesis, we present methods for creating endoscopic video manifolds (EVMs) which allow for more efficient and more robust clustering and classification of endoscopic frames. Given the frames of an endoscopic video, we first compute the EVM representation and then define patient specific endoscopic segments by two different clustering steps performed in the EVM representation. For the classification step, a new endoscopic frame is first projected into the low dimensional representation and then classified as belonging to one of this defined segments. EVM representation respects non-linear pairwise relations between data points (endoscopic frames) while mapping each data point into a low dimensional space ($d \ll w \times h$). Thus, by construction visually different segments become more separated and more compact in this low dimensional representation compared to the original image representation. This allows for a more robust and efficient clustering and classification of endoscopic frames within our re-targeting framework.

In this work, we compute this low dimensional EVM with the following steps:

1. Step 1: Defining the similarities between the data points,
2. Step 2: Constructing the adjacency graph,
3. Step 3 (Optional): Including temporal constraints,
4. Step 4: Computing the Eigenmaps,
5. Step 5: Mapping the data to the low dimensional EVM representation,
6. 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 [Belkin and Niyogi, 2003]. The choice of this particular manifold learning method for creating EVMs is influenced by its strong connections to spectral clustering methods such as [Shi and Malik, 2000]. Laplacian eigenmaps is a local spectral manifold learning technique; i.e. each data point is connected only with its *local* neighbours resulting in a sparse matrix used in the eigenvalue problem. Thus, this method

attempts to preserve only the local neighbourhoods, whereas global methods such as in [Tenenbaum et al., 2000] aim at preserving the geodesic distances between all pairs of data points. This locality-preserving character of the Laplacian eigenmaps intrinsically emphasizes clusters within a dataset [Belkin and Niyogi, 2003] and therefore makes it more suitable for the clustering tasks we address in this work.

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 of informative frames and of different endoscopic segments. In Chapter 7, we demonstrate that EVMs can be adapted to the particular clustering task only by changing the introduced similarity measure. Furthermore, we provide an optional step (Step 3) that allows for including the temporal constraints while also taking the visual similarities between data points into account. Next, we discuss the individual steps for creating EVMs in more detail.

6.1 Defining the Similarities

For each pair $(\mathbf{I}_i, \mathbf{I}_j)$ of the given n data points $\mathbf{I} = \{\mathbf{I}_i\}$, $i, j \in \{1, \cdots, n\}$, first a similarity measure is defined $\mathcal{S} : \mathbf{I} \times \mathbf{I} \rightarrow \mathbb{R}$. $\mathcal{S}$ determines which images are considered to be similar and therefore kept as neighbours 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 respect to the theoretical background discussed in Chapter 5, the chosen similarity measure effects the matrix $\mathbf{H}$ of the eigenvalue problem (Equation 5.1) and thus influences the corresponding operator $\boldsymbol{\mathcal{H}}$ (Equation 5.16) or equivalently the Hilbert space $\mathscr{H}$ (Equation 5.29). At this point, it is important to note that there exist two factors within the spectral manifold learning framework which define the solved eigenvalue problem (Equation 5.1): First the choice of the particular manifold learning technique. As discussed in Chapter 5, various manifold learning techniques preserve different properties of the manifold in the low dimensional space and define the eigenvalue problem accordingly. The second factor is the choice of the similarity measure to define the connections between data points. According to different similarity measures, connections between data points vary and this alters the structure of the approximated manifold. This fact can also be used to include prior knowledge on the relations of data points, as we discuss in Section 6.3. In Chapter 7, we present two similarity measures designed for adapting the manifold structure to two different clustering tasks; i.e. clustering of informative frames and patient specific endoscopic segments.

6.2 Constructing the Adjacency Matrix

Given the similarity matrix S, where the values $S(i, j)$ stand for pairwise similarities between the frames $\mathbf{I}_i$ and $\mathbf{I}_j$ according to the chosen similarity measure $\mathcal{S}(\mathbf{I}_i, \mathbf{I}_j)$, first,

the k-nearest neighbours of each data point are computed. Then, the adjacency matrix is created as:

$$\mathbf{W}(i,j) = \begin{cases} 1 & \text{if } \mathbf{I}_i \in \mathcal{N}_j^{\text{sim}} \\ 0 & \text{otherwise,} \end{cases} \tag{6.1}$$

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

In [Belkin and Niyogi, 2003], the authors propose an optional weighting of the edges of an adjacency graph $\mathbf{W}$ with a heat (Gaussian) kernel for a chosen time parameter t. The weighted adjacency matrix is defined as:

$$\mathbf{W}(i,j) = \begin{cases} e^{-\frac{||\mathbf{I}_i - \mathbf{I}_j||^2}{t}} & \text{if } \mathbf{I}_i \in \mathcal{N}_j^{\text{sim}} \\ 0 & \text{otherwise .} \end{cases} \tag{6.2}$$

The adjacency matrix defined as in Equation (6.1) is referred to as the *combinatorial* adjacency and corresponds to using a Gaussian kernel weighting with $t = \infty$ [Belkin and Niyogi, 2003]. For creating EVMs in this thesis, we rely on the combinatorial adjacency (Equation 6.1). Section 6.7.2 presents a discussion on the motivation behind this choice.

6.3 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 [Atasoy et al., 2010b]. This measure is based on the smoothness of the optical flow field between two temporally related frames and results in the clustering of sequential frames together as long as the optical flow between these frames is a smooth vector field. This temporal constraint is suitable for segmenting the endoscopic video into continuous video segments with smooth endoscope motion. However, it does not allow for clustering temporally distant frames into the same cluster and therefore is not suitable for the clustering of visually similar frames into the same endoscopic segments. Within our clustering and classification framework, we aim at defining meaningful patient specific endoscopic segments by clustering frames showing the same anatomical region together.

To this end, we introduce a method for combining the temporal constraints with the visual similarities simply by defining an additional neighbourhood $\mathcal{N}_j^{\text{temp}}$ based on the temporal order of the frames within the endoscopic video [Atasoy et al., 2011, Atasoy et al., 2012]. In order to account for both, the temporal relations as well as visual similarities, the adjacency matrix is defined as:

$$\mathbf{W}(i,j) = \begin{cases} 1 & \text{if } \left((\mathbf{I}_i \in \mathcal{N}_j^{\text{sim}}) \text{ or } (\mathbf{I}_i \in \mathcal{N}_j^{\text{temp}}) \right) \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

In our experiments, we use the same number k for defining the visual $\mathcal{N}_j^{\text{sim}}$ and the temporal $\mathcal{N}_j^{\text{temp}}$ neighbourhoods. 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 neighbourhood 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 thesis, we refer to the EVMs with temporal constraints as visual and temporal endoscopic video manifolds (vtEVMs).

6.4 Computing Laplacian Eigenmaps

After creating the adjacency graph, the graph Laplacian $\mathbf{L}$ is computed as:

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \ , \tag{6.4}$$

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

$$\mathbf{D}(i,i) = \sum_{j=1}^{n} \mathbf{W}(i,j) \ . \tag{6.5}$$

The Laplacian eigenmaps are determined as the eigenvectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ of the Laplacian matrix $\mathbf{L}$ corresponding to the d smallest non-zero eigenvalues $0 < \lambda_1 \leq \cdots \leq \lambda_d$. As each of these Laplacian eigenmaps $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$, with d being the dimensionality of the manifold, solves the generalized eigenvalue problem $\mathbf{L}\mathbf{v} = \lambda\mathbf{D}\mathbf{v}$, they minimize the following objective function [Belkin and Niyogi, 2003]:

$$\sum_{i,j \in \{1,\cdots,n\}} (\mathbf{y_i} - \mathbf{y_j})^2 \mathbf{W}(i,j) \ , \tag{6.6}$$

where $\mathbf{y_i} = [\mathbf{v}_1(i), \cdots, \mathbf{v}_d(i)]^\top$ denotes the d-dimensional representation of the data point $\mathbf{I}_i$.

6.5 Endoscopic Video Manifold (EVM) Representation

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

$$\mathbf{y_i} = [\mathbf{v}_1(i), \cdots, \mathbf{v}_d(i)]^\top \ . \tag{6.7}$$

The eigenvectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ are the discrete equivalents of the eigenfunctions of the Laplace-Beltrami operator in the continuous domain as applied onto the manifold as discussed in Section 5.5.2. Note that using the Laplacian Eigenmaps [Belkin and Niyogi, 2003] 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 Section 6.6.

6.6 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 [Beyer et al., 1999, He et al., 2005]. Although the dimensionality can be reduced as described in the previous steps with non-linear manifold learning, these techniques require all data points to be available to compute their low dimensional representation. Unlike linear methods such as PCA, non-linear manifold learning techniques compute the projections of the data points onto the low dimensional representation *without explicitly estimating the non-linear function to map the data from the high to the low dimensional space*; i.e. the projection of a data point $\mathbf{I}_i$ into the low dimensional space is computed as $\mathbf{y}_i = \nu(\mathbf{I}_i)$, $\mathbf{I}_i \in \mathbf{I}$ without explicitly estimating the continuous non-linear 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 $\mathbf{I}_s \notin \mathbf{I}$ cannot be easily computed by evaluating the mapping function $\nu(\mathbf{I}_s) = \mathbf{y}_s$ for an unknown data point $\mathbf{I}_s$.

The Locality Preserving Projections (LPP) method [He et al., 2005] estimates the optimal *linear* mapping from the high to low dimensional representations by optimizing the same objective function as defined in the Laplacian eigenmaps method (Equation 4.19). 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 $\mathbf{y}_i^{(1)}$ of a high dimensional data point $\mathbf{I}_i$ onto the one dimensional line be defined using a linear mapping:

$$\mathbf{y}_i^{(1)} = \widetilde{\nu}(\mathbf{I}_i) = \mathbf{t}^\top \mathbf{I}_i \ , \tag{6.8}$$

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

Substituting Equation (6.8) in the objective function given in Equation (6.6) results in:

$$\begin{aligned} &\sum_{i,j} \left(\mathbf{y}_i^{(1)} - \mathbf{y}_j^{(1)} \right)^2 \mathbf{W}(i,j) = \\ &\sum_{i,j} \left(\mathbf{t}^\top \mathbf{I}_i - \mathbf{t}^\top \mathbf{I}_j \right)^2 \mathbf{W}(i,j) \ , \end{aligned} \tag{6.9}$$

which in turn can be expressed in matrical form:

$$\sum_{i,j} \left(\mathbf{y}_i^{(1)} - \mathbf{y}_j^{(1)} \right)^2 \mathbf{W}(i,j) = \mathbf{t}^\top \mathbf{I} \mathbf{L} \mathbf{I}^\top \ , \tag{6.10}$$

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

$$\mathbf{I} \mathbf{L} \mathbf{I}^\top \mathbf{t} = \lambda \mathbf{I} \mathbf{D} \mathbf{I}^\top \mathbf{t} \ . \tag{6.11}$$

This solution can be extended to higher dimensional cases simply by defining the linear mapping $\mathbf{T} : \mathcal{M} \mapsto \mathbb{R}^d$ as $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_d]$. For the detailed derivation of Equations (6.10) and (6.11) from Equation (6.9) we refer to [He et al., 2005].

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 $\mathbf{T}$:

$$\widetilde{\nu}(\mathbf{I}_s) = \mathbf{T}^\top \mathbf{I}_s = \mathbf{y}_s \ , \text{ with } \ ,[\mathbf{t}_1, \cdots, \mathbf{t}_d] \ , \tag{6.12}$$

where $\mathbf{T}$ denotes the linear projection matrix and provides a function basis containing the eigenvectors $[\mathbf{t}_1, \cdots, \mathbf{t}_d]$ as explained in [He et al., 2005].

6.7 EVM Parameters

While creating the EVM representation, three parameters can be chosen according to the addressed task and application. Firstly, the number of manifold neighbours, i.e. number of connections per data point in the adjacency graph, should be determined considering the shared information and relations between the individual data points. Secondly, the weightings used in the adjacency graph have to be defined. Finally, the manifold dimensionality needs to be chosen by the user for datasets such as endoscopic videos, on which the graph Laplacian exhibits a nearly continuous spectrum.

Furthermore, in this thesis we extend the classical framework for non-linear manifold learning by introducing new similarity measures. These introduced similarity measures do not necessarily induce further parameters in the classical framework, but they permit the flexibility of adapting the manifold structure according to the addressed clustering task. Examples for selecting or designing the similarity measure in accordance with the particular clustering task are presented in Sections 7.2.1 and 7.3.1. As next we present some discussion for the effect of manifold learning parameters on defining meaningful representations of endoscopic datasets.

6.7.1 Manifold Neighbourhoods

The number of NN (k) used in creating the manifold neighbourhood regulates the connectedness of the adjacency graph and is in general estimated empirically. As explained in Section 6.2, EVMs rely on k-NN of each data point while creating the adjacency graph. This means that each point will only be connected to its k-NN according to the chosen similarity measure S. The choice of the parameter k depends on two factors; relations between data points, more precisely on the shared information content between endoscopic frames and the target application.

In this thesis, we address two different clustering tasks: first clustering of informative/uninformative frames, and second clustering of different endoscopic segments. Intuitively, the choice of EVM parameters for the first clustering task is not trivial.

Therefore, in Section 7.2.3, we provide a quantitative evaluation for the effect of this parameter on the clustering accuracy.

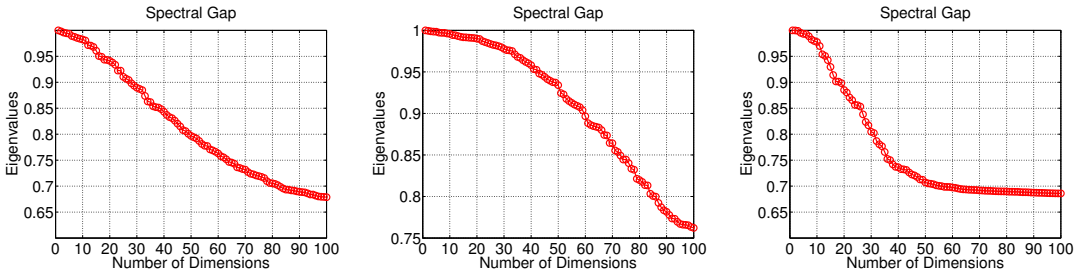
The aim of the second clustering step is to create clusters of endoscopic frames in accordance with the anatomical region in their content; i.e. grouping frames showing the same anatomical regions together while assigning different clusters to different anatomical regions. Thus, in the EVM representation, frames showing the same anatomy should be located close to each other whereas frames acquired at different locations are more separated. To achieve this, we chose the value of k according to the number of frames showing the same anatomical region. Considering the frame rate in a standard endoscopic acquisition (25 frames per second) and the endoscope motion, we choose $k = 20$ in our experiments for clustering endoscopic segments. For all endoscopic datasets used in our study, this value results in one connected component in the adjacency graph and yields visually meaningful endoscopic segments. Clusters defined on EVMs created with $k = 20$ are illustrated in Figure 7.11 in Chapter 7.

6.7.2 Weightings of the Adjacency Matrix

A manifold created with the combinatorial adjacency matrix (Equation 6.1) 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 is lower than the similarity of two 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 is the same as of the to their consecutive frames. Thus, in the low dimensional representation, frames with slightly different endoscope viewpoints will also be closely localized and clustered into the same endoscopic segment. Because of this desired property, we create EVMs using the combinatorial adjacency graph.

6.7.3 Manifold Dimensionality

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 non-linear 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 $\mathbf{L}$ exhibits a near to continuous spectrum and does not provide a clear estimation of dimensionality. The spectrum of the graph Laplacian $\mathbf{L}$ computed on three upper GI endoscopic datasets is illustrated in Figure 6.1. Due to the absence of a significant spectral gap to reveal the dimensionality, in Sections 6.7.3 and 6.7.3, we provide an evaluation on the effect of the parameter d for clustering uninformative



(a) (b) (c)

Figure 6.1: Spectrum of the graph Laplacian **L** computed on three upper GI endoscopic datasets. There exist no clear gap in the spectrum to be used for determining the dimensionality of the EVM representation.

frames and endoscopic segments, respectively. In Sections 7.2.3.1 and 7.3.3.1, we quantitatively evaluate the manifold dimensionality with respect to the efficiency of clustering informative frames and endoscopic segments.

Part III

Targeted Optical Biopsies on Endoscopic Video Manifolds

Chapter 7

Clustering of Diagnostic Endoscopy

> 'It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.'
>
> Richard P. Feynman

In this thesis, we formulate the intra-video localisation task for optical biopsy re-targeting as a clustering and classification problem. As first, in the offline processing stage, we *define patient specific endoscopic segments (PSESs)* by a two step clustering of the diagnostic endoscopy. In the second stage, the online classification, these PSESs are used as classes (labels) and to each frame of the surveillance endoscopy a label is assigned during the examination. If a frame is assigned an optical biopsy class, then this location is recognized as a previous optical biopsy site and the endoscopic expert is notified. Both clustering and classification parts are performed in the EVM representation while adapting the structure of the manifold according to the addressed task.

In this Chapter, we first give an overview of the offline processing stage and then explain each of the two clustering steps individually. We discuss two similarity measures introduced to adapt the EVM structure according to each clustering step. Finally, we present quantitative evaluation for both steps of the offline clustering.

7.1 Overview of the Offline-Processing

A typical endoscopic video contains around $40\% - 50\%$ uninformative frames. This is due to the particular imaging conditions encountered in upper GI endoscopy as explained in Section 2.3.1. Furthermore, there exist no temporal relation between informative and uninformative frames as illustrated in Figure 7.1.

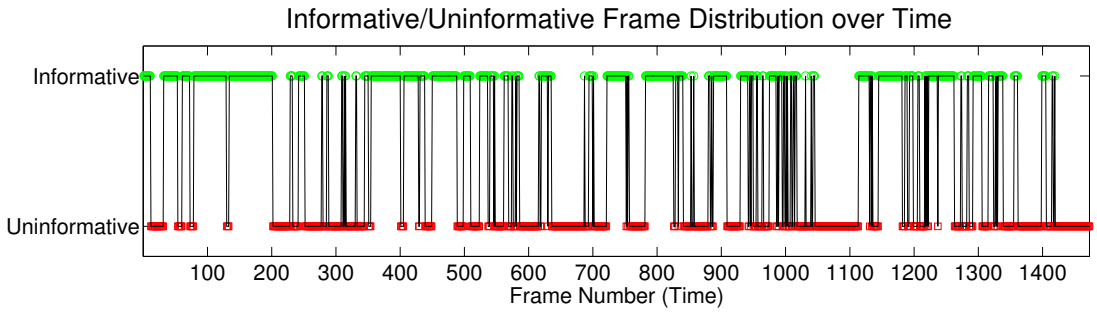


Figure 7.1: Distribution of informative and uninformative frames over time. Green circles mark the informative frames whereas red squares show the uninformative frames labelled by an endoscopic expert. Black line illustrates the temporal continuity of the video. Frequent switching between the informative and uninformative frames, as seen on the black line, suggests the poor temporal relation within the informative and the uninformative frames.

In the offline processing of the diagnostic dataset, endoscopic frames are first clustered according to their informativeness (image quality). To this end, a corresponding EVM representation is created using a new similarity measure, i.e. the energy histogram similarity and the clustering is performed in this low dimensional representation (Figure 7.2(a), 7.2(b)). From the results of this unsupervised clustering, informative clusters are selected by the endoscopic expert (Figure 7.2(c)). This way, the decision on the usefulness of the frame clusters is made by an endoscopic expert while the grouping of the individual frames into clusters is performed in an unsupervised manner based on the image quality. The informative clusters selected by the endoscopic expert are used to define the PSESs. For this, a second clustering is performed to group these informative frames this time according to their visual content (Figure 7.2(d)-7.2(f)). The goal of this second clustering step is to combine frames showing the same scene into the same cluster, where each cluster defines one PSES. Steps of the offline processing stage are illustrated in Figure 7.2.

7.2 Clustering Informative Frames

The first step of our proposed framework consists of clustering uninformative frames on the accordingly created EVM. Given a diagnostic endoscopic video, we first create an EVM representation that allows the clustering of uninformative frames. In order to adapt the manifold structure to this particular clustering task, a novel similarity measure (energy histogram similarity) is introduced, which was first explored in our previous study [Atasoy et al., 2010b]. 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.

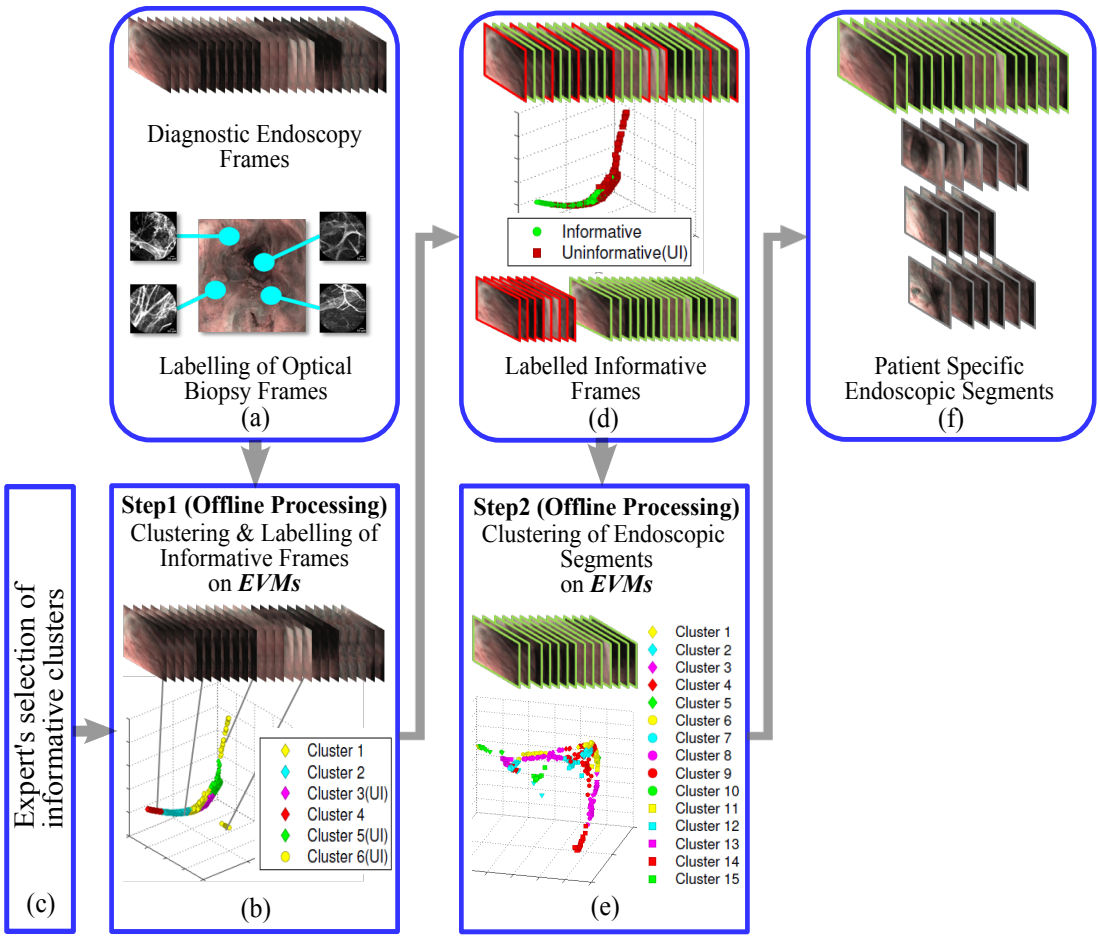


Figure 7.2: Offline processing stage of the proposed clustering-classification framework. (a) Input of the offline processing step consists of the frames of the diagnostic endoscopy and labelling of the frames where an optical biopsy has been acquired. (b) Clustering of the informative and uninformative 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.

7.2.1 Energy Histograms Kernel

In order to create an EVM, where the uninformative frames are closely localized, the similarity measure used for manifold learning (Section 6.1) needs to be designed such that it yields a low similarity between informative and uninformative frames and a high similarity for two informative or two uninformative frames. To achieve this, 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 compared to an uninformative frame whose energy mainly accumulates in the low frequencies (Figure 7.3). To create a measure of similarity between the power spectra, first the Fourier transform f_i of an endoscopic frame $\mathbf{I}_i$ is computed and represented in log-polar coordinates $f_i(\omega, \theta)$, where $\omega \in [\omega_0, \cdots, \omega_x, \cdots, \omega_r]$ is the set of considered frequencies and $\theta \in [0, \cdots, 2\pi]$ is the discrete set of orientations. In order to achieve rotation invariance, the 2D spectrum is integrated over θ resulting in an $r \times 1$ dimensional vector $\mathcal{F}_i$, r being the number of different frequencies of the discrete Fourier transform:

$$\boldsymbol{\mathcal{F}}_i(\omega) = \sum_{\theta=0}^{2\pi} f_i(\omega, \theta) \quad . \tag{7.1}$$

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

$$\gamma_\delta = \left\{ \omega_x \middle| \frac{\delta - 1}{b} \leq \omega_x \leq \frac{\delta}{b} \right\} \quad , \quad \forall \delta \in \{1, \cdots, n\} \ , \tag{7.2}$$

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

$$\begin{aligned} \mathbf{h}_i &= [h_{\gamma_1}(\mathcal{F}_i), \cdots, h_{\gamma_b}(\mathcal{F}_i)]^\top \\ h_{\gamma_\delta}(\mathcal{F}_i) &= \sum_{\omega_x \in \gamma_\delta} \mathcal{F}_i(\omega_x) \ . \end{aligned} \tag{7.3}$$

In Section 7.2.3.1, 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:

$$\begin{aligned} S(\mathbf{I}_i, \mathbf{I}_j) &= S_{\text{EH}}(\mathbf{I}_i, \mathbf{I}_j) \\ S_{\text{EH}}(\mathbf{I}_i, \mathbf{I}_j) &= 1 - \left(\frac{\langle \mathbf{h}_i, \mathbf{h}_j \rangle}{||\mathbf{h}_j|| \cdot ||\mathbf{h}_j||} \right) \quad , \end{aligned} \tag{7.4}$$

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 $\mathbf{I}_i$ and uninformative

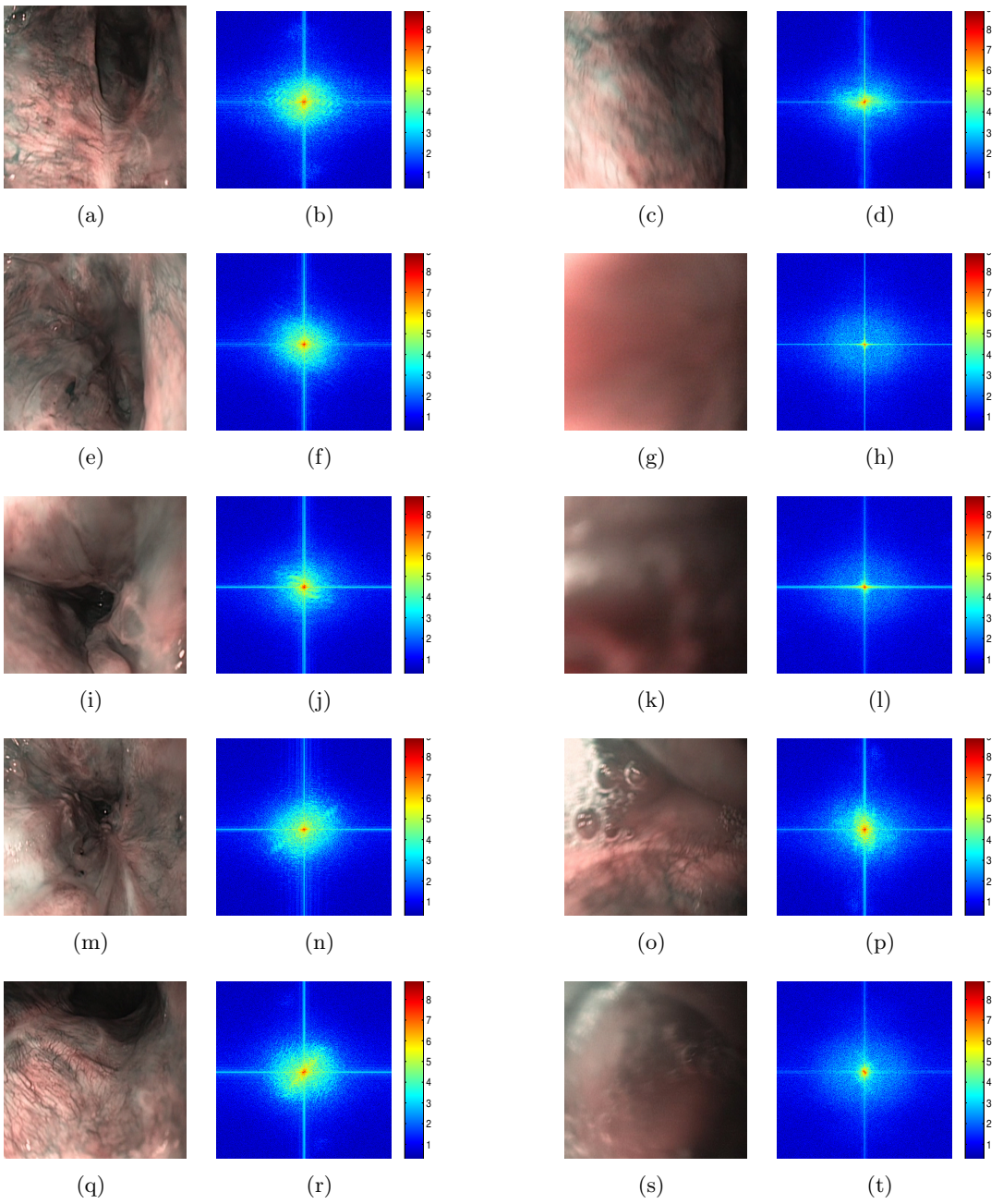


Figure 7.3: (a), (e), (i), (m) and (q) show endoscopic frames in ideal conditions acquired by a state-of-the art GI endoscope. (b), (f), (j), (n) and (r) illustrate the power spectrum of the ideal frames. (c), (g), (k), (o) and (s) show examples of uninformative endoscopic frames. (d), (h), (l), (p) and (t) illustrate the power spectrum of these informative frames. For the clarity of the visualisation, 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. Warmer colours indicate higher values.

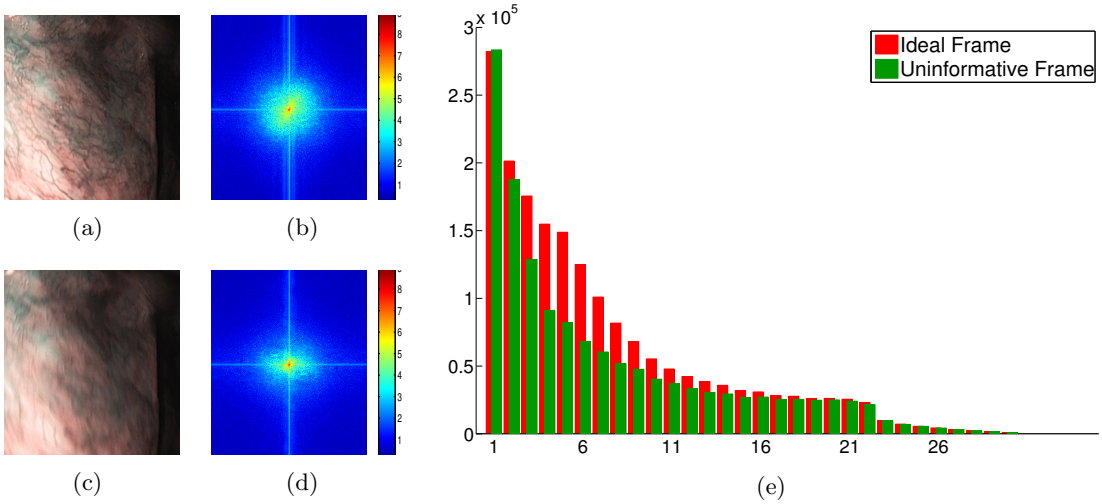


Figure 7.4: (a) An ideal (informative) frame acquired by a state-of-the art GI endoscope. (b) The power spectrum of the informative frame. (c) An endoscopic frame with motion blur. (d) The power spectrum of the motion blurred frame. (e) Rotation invariant energy histograms of an informative and uninformative frame computed using 30 bins.

frame $\mathbf{I}_j$ as shown in Figure 7.4, the energy histogram similarity measure $S_{\text{EH}}(\mathbf{I}_i, \mathbf{I}_j)$ 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 informative and uninformative frames are better separated as illustrated in Figure 7.5.

7.2.2 Clustering on the EVMs

After evaluating the energy histogram similarity measure S_{EH} on all pairs of frames, EVMs are computed as described in Chapter 6. Then, a K-means clustering [Hartigan and Wong, 1979] is performed in the EVM representation. Cluster centres are initialized randomly and 100 trials are performed to ensure a stable clustering. Finally, the resultant clusters $\{C_1^{\text{EH}}, \cdots, C_K^{\text{EH}}\}$ 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 [Von Luxburg, 2007]. The method presented in this thesis 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. This allows us to adapt the EVMs according to different clustering tasks.

7.2.3 Evaluation

In this Section, we present experiments to demonstrate the effectiveness of the informative frame clustering. We evaluate the agreement between the clusters obtained on the EVMs (EVM_{EH}) created with the energy histogram similarity $\mathcal{S}_{\text{EH}}$, and a ground truth labelling. To this end, we compute the recall and precision of the uninformative frame labelling with respect to the parameters of the created EVMs.

Uninformative frames of each dataset are clustered independently on the EVM_{EH} created using the energy histogram S_{EH} similarity matrix as explained in Section 7.2. In the proposed workflow, the resultant clustering is presented to the endoscopic expert during post-processing 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).

The proposed S_{EH} similarity yields well structured manifolds, where informative and uninformative frames are well separated as shown in Figure 7.5(a), 7.5(c) and 7.5(e). An example of the results with 6 clusters is illustrated in Figure 7.5(b), 7.5(d) and 7.5(f).

For quantitative analysis, *recall* and *precision* and *F-measure* quality measures of each clustering are evaluated over a varying number of clusters from 2 to 50. Given a clustering $\mathbf{C}_\kappa^{\text{EH}} = \{C_1^{\text{EH}}, \cdots, C_\kappa^{\text{EH}}\}$ 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. Recall, precision and F-measure are computed as follows:

$$\text{Recall}(\mathbf{C}_\kappa^{\text{EH}}) = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} , \tag{7.5}$$

$$\text{Precision}(\mathbf{C}_\kappa^{\text{EH}}) = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} , \tag{7.6}$$

$$\text{F}(\mathbf{C}_\kappa^{\text{EH}}) = \frac{2 \cdot \text{Precision}(\mathbf{C}_\kappa^{\text{EH}}) \cdot \text{Recall}(\mathbf{C}_\kappa^{\text{EH}})}{\text{Precision}(\mathbf{C}_\kappa^{\text{EH}}) + \text{Recall}(\mathbf{C}_\kappa^{\text{EH}})} . \tag{7.7}$$

7.2.3.1 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 neighbours $k \in [4, \cdots, 20]$, manifold dimensions $d \in [2, \cdots, 20]$ and number

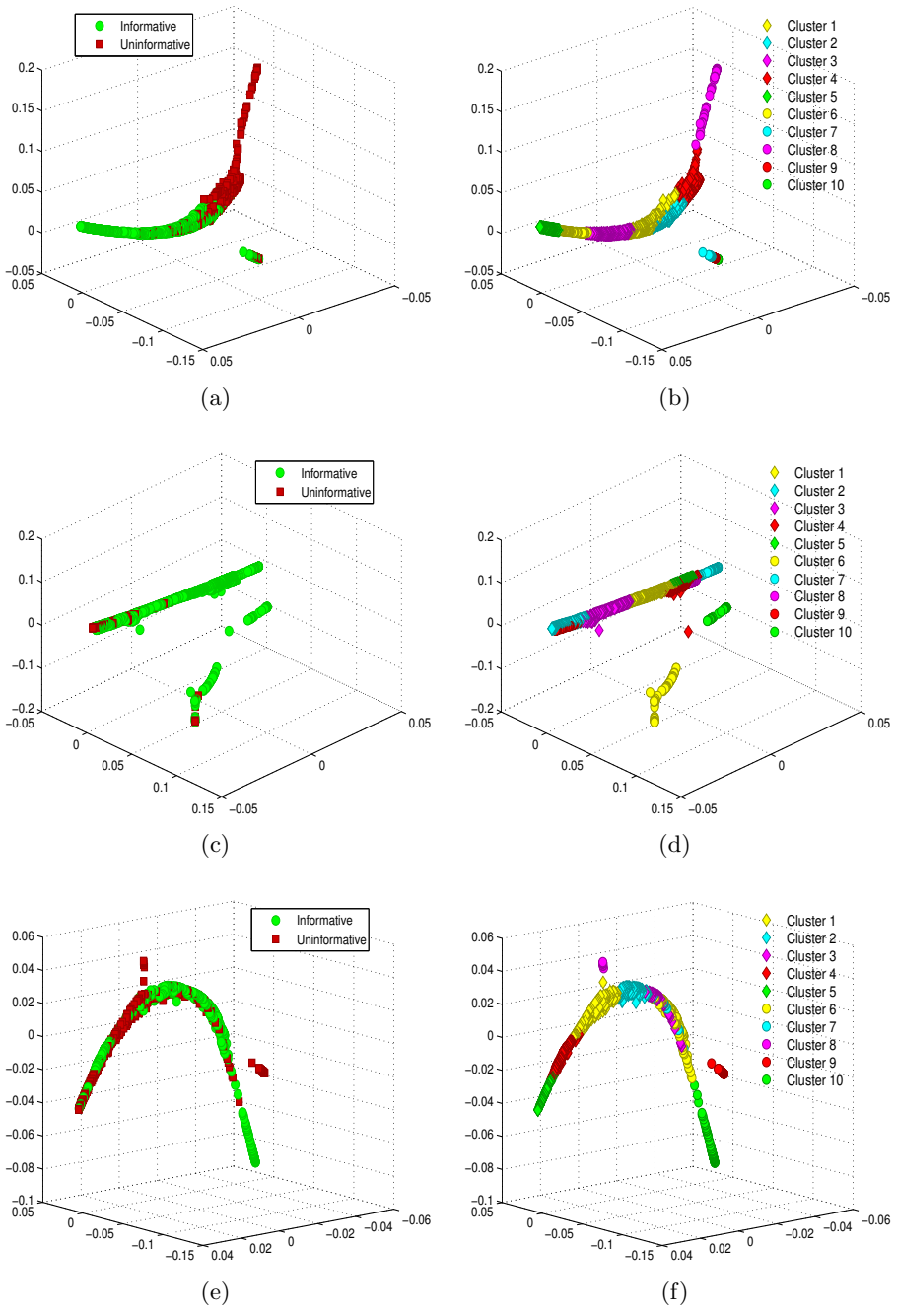


Figure 7.5: (a), (c) and (e) show the 3 dimensional EVMs of 1st, 2nd, and 3rd endoscopic video computed with the proposed $\mathcal{S}_{\text{EH}}$ similarity measure, respectively. The red squares illustrate the uninformative frames in the ground truth labelling. (b), (d) and (f) show the clustering results on the EVMs for the 1st, 2nd, and 3rd endoscopic video, respectively. The use of $\mathcal{S}_{\text{EH}}$ in manifold learning leads to structured EVMs where the uninformative frames are clustered together.

of histogram bins used in the energy histograms $b \in [10, \cdots, 160]$ while changing the number of clusters from 2 to 50 $\kappa \in [2, \cdots, 50]$. Figure 7.6 illustrates F-measures of uninformative frame labelling on EVMs created with all evaluated parameter values. The F-measure combines the precision and recall values of the labelling into one quality measure and takes values in $[0, 1]$. For the best recall and precision 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. This 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 Figures 7.6(g), 7.6(h) and 7.6(i), 30 histogram bins result in near to optimum F-measure values for all three datasets.

Recall-precision plots of all clustering results computed on EVMs with dimensionality $d = 3$, $d = 6$ and $d = 9$ are demonstrated in Figure 7.7(a)-7.7(c). Similarly, the recall-precision values estimated from EVMs with $k = 6$, $k = 10$ and $k = 14$ manifold nearest neighbours are shown in Figure 7.7(d)-7.7(f). The large overlap of recall-precision curves estimated from EVMs with different dimensionality and different neighbourhoods 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. Figure 7.7(g)-7.7(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 recall-precision values.

Finally, we also evaluate the threshold that is used for determining the ground truth informative/uninformative clusters. Figure 7.8 shows that if the threshold is smaller than 50% the evaluation becomes more sensitiv to the presence of the informative frames within a cluster but its precision 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.

7.2.3.2 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 neighbours $k \in [4, \cdots, 20]$ and dimensionality $d \in [2, \cdots, 20]$ 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 discretisation of the energy histogram (EH) to a more advanced estimation of the energy distribution (EDstr). By approximating the energy distributions with Parzen windowing, we use 30 samples

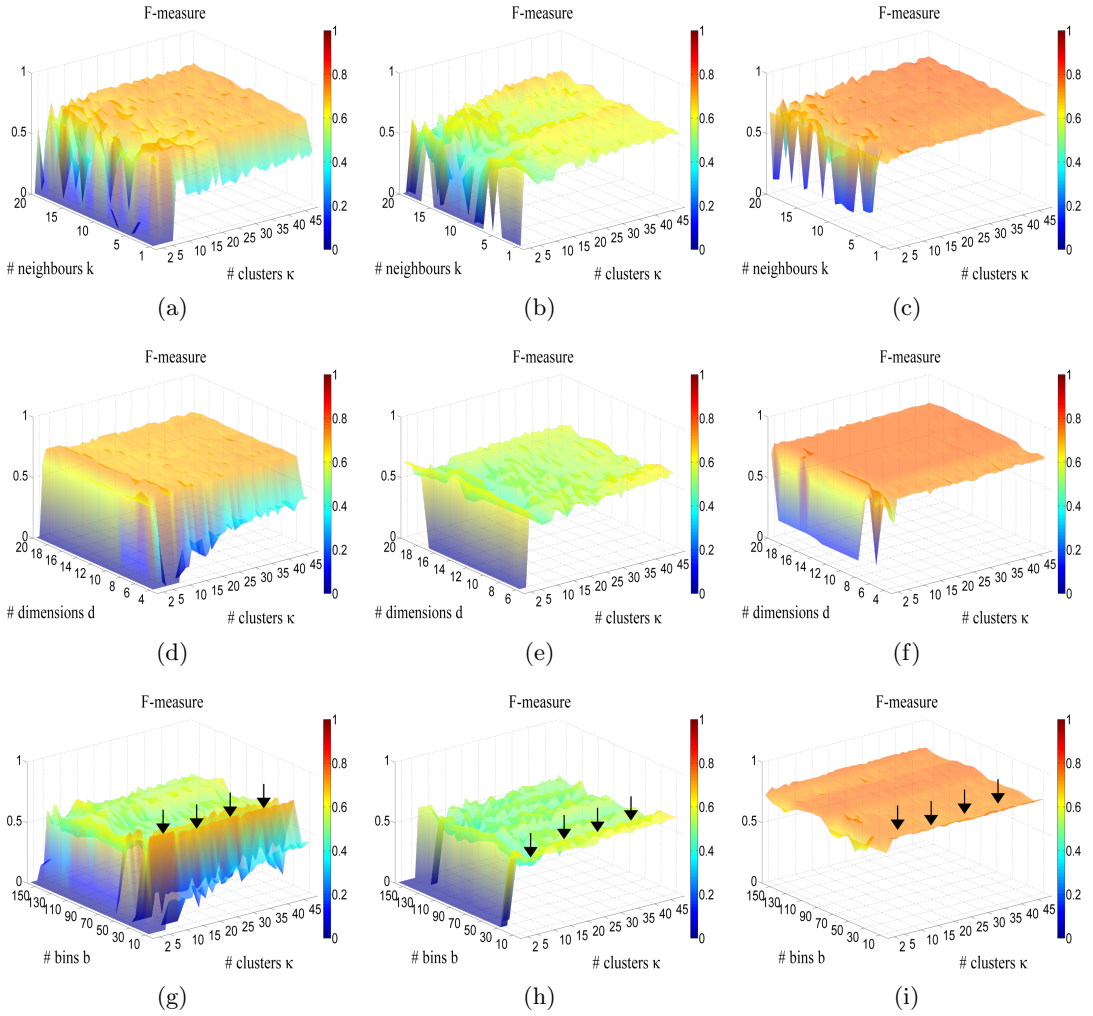


Figure 7.6: Evaluation of the EVM parameters for clustering uninformative frames. Recall, precision and F-measure values are evaluated for different number of manifold nearest neighbours k and dimensions d while changing the number of clusters. (a), (b) and (c) show the F-measure values as surface-plots for d ranging from 1 to 20 with $k = 6$ and $b = 30$ for the 1st, 2nd and 3rd dataset respectively. (d), (e) and (f) show the F-measure values for number of manifold neighbours $k \in [4, \cdots, 20]$ 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 presented clustering to different parameter values. (g), (h) and (i) illustrate the F-measure values for different number of bins $b \in [10, \cdots, 160]$ used in energy histograms and the number of clusters $\kappa \in [2, \cdots, 50]$ 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 our experiments.

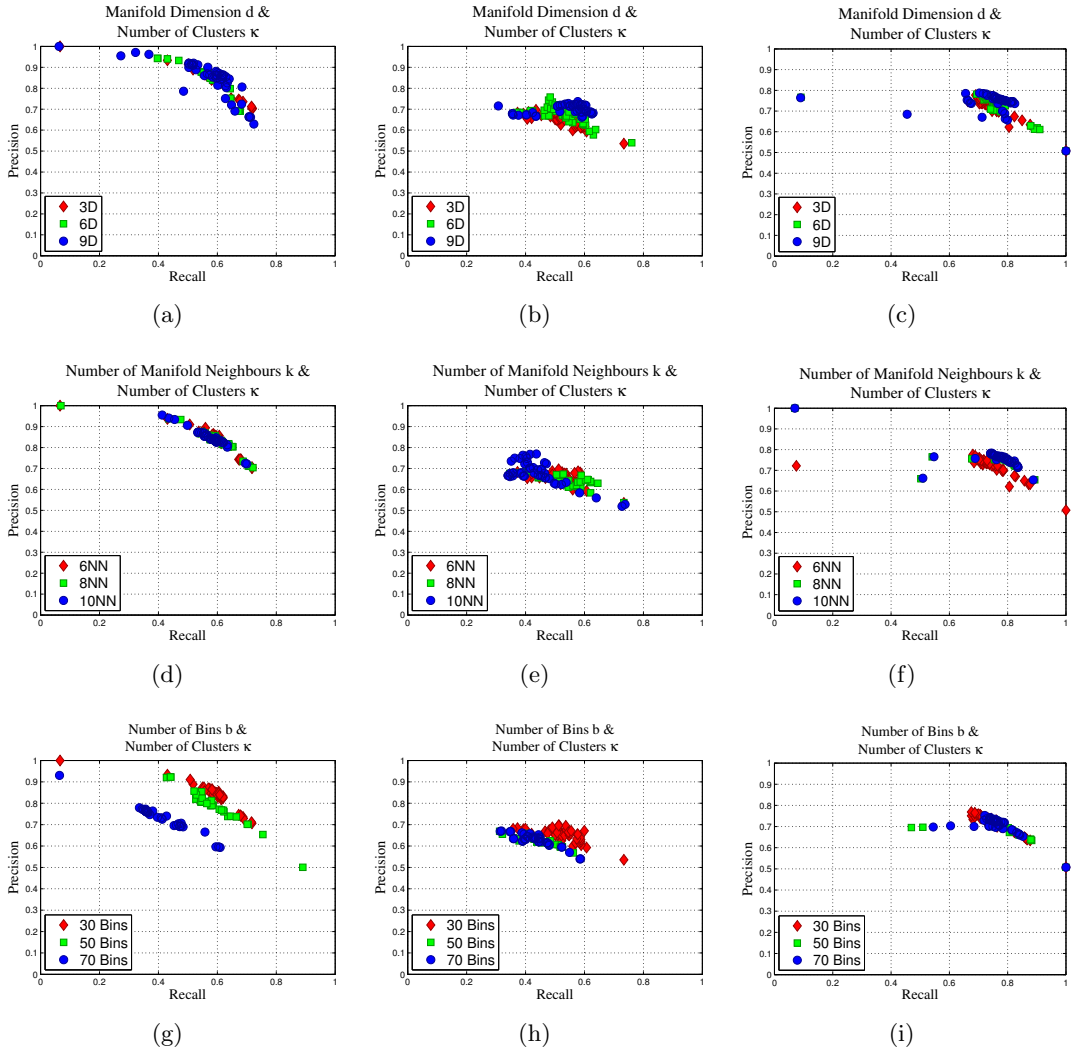
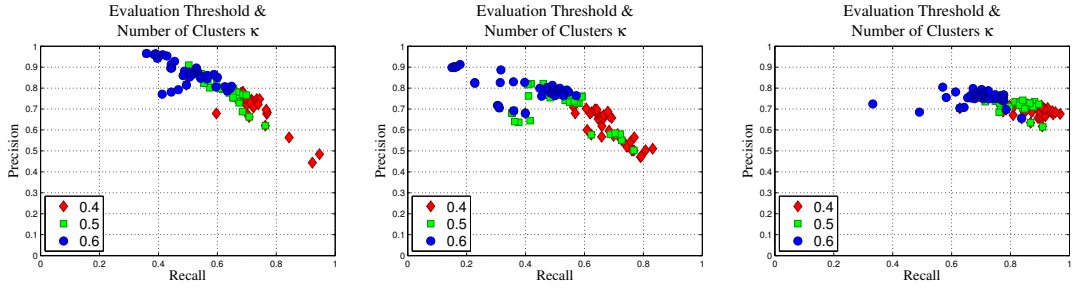


Figure 7.7: a), (b) and (c) show the recall versus precision plots of the clusterings with number of clusters κ ranging between 2 and 50, where the clusterings are performed on EVM_{EH} with $d = 3$, $d = 6$ and $d = 9$ ($b = 30$ and $k = 6$) computed using the S_{EH} similarity matrix from the 1st, 2nd and 3rd dataset, respectively. (d), (e) and (f) show the recall-precision 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 neighbours ($d = 3$ and $b = 30$) from the 1st, 2nd and 3rd dataset, respectively. (g), (h) and (i) display the recall versus precision plots for the 1st, 2nd and 3rd 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 $d = 3$.



(a) (b) (c)

Figure 7.8: (a), (b) and (c) show the recall-precision 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.

(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 discretisation into 30 bins). Table 7.1 summarizes the performance of all 4 methods; *i.e.* cosine and Bhattacharyya distances on EH and EDstr, on the 3 datasets.

We have observed that the cosine measure used in S_{EH} in Equation (7.4) separates the informative and uninformative frames better compared to the Bhattacharyya distance. Furthermore, energy histograms as defined in Equation (7.3) lead to higher F-measures compared to smoother energy distributions. A possible explanation is that a simple discretisation 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.

7.3 Defining the Patient Specific Endoscopic Segments (PSESs)

In the second step of our framework, we define patient specific endoscopic segments by clustering the labelled informative frames of the diagnostic endoscopy. To this end, we create a new EVM representation by introducing another similarity measure. 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. Therefore, we use the

		EH **30 bins**	***ED*** **30 samples**	***ED*** **100 samples**	***ED*** **448 samples**
Number of Manifold Neighbours					
Data 1	***Cos.***	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)
	Bhat.	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	***Cos.***	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)
	Bhat.	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	***Cos.***	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)
	Bhat.	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	***Cos.***	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)
	Bhat.	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	***Cos.***	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)
	Bhat.	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	***Cos.***	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)
	Bhat.	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)

Table 7.1: F-measure values of Cosine (Cos.) and Bhattacharyya (Bhat.) distance measures evaluated on energy histograms (EH) and energy distributions (EDstr) computed using Parzen windowing on the 3 datasets. EVMs for both measures; i.e. Cosine (Cos.) and Bhattacharyya (Bhat.) distances are created by ranging the parameters $k \in [6, \cdots, 20]$, $d \in [1, \cdots, 20]$ and $\kappa \in [2, \cdots, 50]$.

normalized cross correlation (NCC) similarity measure to define the neighbourhood in the manifold learning framework.

7.3.1 NCC Similarity Measure

In the original setting of the Laplacian Eigenmaps method, involved in computing the EVMs, the standard choice of the distance (dis-similarity) measure is the Euclidean distance [Belkin and Niyogi, 2003] 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:

$$\mathrm{NCC}(\mathbf{I}_i, \mathbf{I}_j) = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \frac{(\mathbf{I}_i(x,y) - \overline{\mathbf{I}_i})(\mathbf{I}_j(x,y) - \overline{\mathbf{I}_j})}{\sigma_{\mathbf{I}_i} \sigma_{\mathbf{I}_j}} \ , \tag{7.8}$$

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

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

$$\mathrm{NCC}(\mathbf{I}_i, \mathbf{I}_j) = \langle \phi_{\mathrm{NCC}}[\mathbf{I}_i], \phi_{\mathrm{NCC}}[\mathbf{I}_j] \rangle \tag{7.9}$$

with

$$\phi_{\mathrm{NCC}}[\mathbf{I}_i](x,y) = \frac{\mathbf{I}_i(x,y) - \overline{\mathbf{I}_i}}{||\mathbf{I}_i(x,y) - \overline{\mathbf{I}_i}||} \ . \tag{7.10}$$

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 measure inducing a topology on the dataset.

7.3.2 Clustering on the EVMs

We include the NCC similarity measure into the EVM framework by choosing the similarity measure $S(\mathbf{I}_i, \mathbf{I}_j)$, as explained in Section 6.1, to be equal to the NCC measure $S_{\mathrm{NCC}}\ (\mathbf{I}_i, \mathbf{I}_j) = \mathrm{NCC}(\mathbf{I}_i, \mathbf{I}_j)$. Using the S_{NCC} we compute the new low dimensional representation of the informative frames, i.e. $\mathrm{EVM}_{\mathrm{NCC}}$. Then the K-means clustering [Hartigan and Wong, 1979] is performed in this representation and each cluster in $\{C_1^{\mathrm{PSES}}, \cdots, C_\beta^{\mathrm{PSES}}\}$ is defined as one PSES.

7.3.3 Evaluation

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 the original image representation and to principal component analysis is presented.

7.3.3.1 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) [Davies and Bouldin, 1979] for each clustering while varying the manifold dimensionality from 1 to 40 and the number of clusters from 2 to 40.

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

$$\begin{aligned} \mathrm{WCD}(\mathbf{C}_\kappa) =& [\mathrm{WCD}(C_1), \cdots, \mathrm{WCD}(C_\kappa)]^\top , \\ \mathrm{WCD}(C_i) =& \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \overline{\mathbf{z_i}}|| , \end{aligned} \tag{7.11}$$

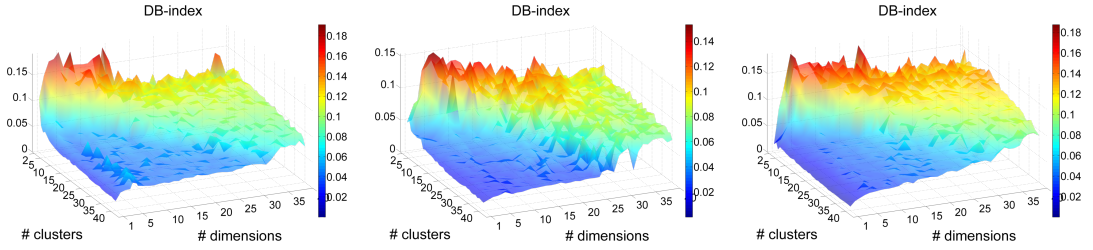
where $\overline{\mathbf{z_i}}$ denotes the centre of cluster C_i and n_i denotes the number of elements in cluster C_i. WCD indicates the compactness of the clusters (the smaller the WCDs, the more compact the clusters) and is evaluated in relation to the separability of the clusters which is measured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

$$\begin{aligned} \mathrm{BCD}(\mathbf{C}_\kappa) =& [\mathrm{BCD}(C_1), \cdots, \mathrm{BCD}(C_\kappa)]^\top , \\ \mathrm{BCD}(C_i) =& \sum_{j=1, j \neq i}^{\kappa} ||\overline{\mathbf{z_i}} - \overline{\mathbf{z_j}}|| . \end{aligned} \tag{7.12}$$

The DB-index is computed as:

$$\mathrm{DB}(\mathbf{C}_\kappa) = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \sum_{j=1, j \neq i}^{\kappa} \max \left(\frac{\mathrm{WCD}(C_i) + \mathrm{WCD}(C_j)}{||\overline{\mathbf{z_i}} - \overline{\mathbf{z_j}}||} \right) . \tag{7.13}$$

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 analysed and its value only depends on the appropriateness of the clustering, which is related to the actual number of clusters in the dataset [Davies and Bouldin, 1979]. 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.



(a) (b) (c)

Figure 7.9: Evaluation of manifold dimensionality d in relation to the number of clusters κ. DB-indices are evaluated by varying the number of clusters κ from 2 to 40 and the manifold dimensionality d from 1 to 40 on the (a) 1st, (b) 2nd, and (c) 3rd dataset. DB-indices show a significant decrease once the number of clusters is equal or greater the manifold dimensionality.

The DB-indices for the 1st, 2nd and 3rd datasets are shown in Figures 7.9(a), 7.9(b) and 7.9(c), respectively. In agreement with the study in [Von Luxburg, 2007], 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 Figure 7.9. In the remaining of the experiments, we choose the number of clusters to be 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.

7.3.3.2 Comparison of Data Representations

To demonstrate that the EVM representation is better suited for defining PSESs, we conduct experiments comparing clusterings in the original image representation, its principal components and different 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 computed based on the within- and between-cluster distances and second the DB-index (Equation 7.13) is evaluated to measure the separation between clusters. EVM clusterings are compared to K-means clustering performed on the original images and on the linear manifold computed using principal component analysis (PCA). The six compared data representations will be referred to as follows in the rest of this thesis:

- ImSp: original grey 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 [Fukunaga and Olsen, 1971],

- EVM_{ED}: non-linear manifold computed using Laplacian Eigenmaps with the standard Euclidean distance measure,
- EVM_{NCC}: non-linear manifold computed using Laplacian Eigenmaps with the NCC measure,
- vtEVM_{ED}: visual-temporal manifold computed using Laplacian Eigenmaps with the standard Euclidean distance measure including the temporal constraints as described in Section 6.3,
- $\text{vtEVM}_{\text{NCC}}$: visual-temporal manifold computed using Laplacian Eigenmaps with the NCC measure including the temporal constraints as described in Section 6.3.

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

$$\text{CS}(\mathbf{C}_\kappa) = \frac{\min(\text{BCD}(\mathbf{C}_{\boldsymbol{\kappa}}))}{\frac{1}{\kappa}\sum_{c=1}^{\kappa}\text{WCD}(C_c)} \quad . \tag{7.14}$$

Figures 7.10(a), 7.10(c) and 7.10(e) show 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 neighbours 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 [Davies and Bouldin, 1979]. Figures 7.10(b), 7.10(d) and 7.10(f) show the evaluation of DB-index for different number of clusters ranging from 5 to 50 for EVM_{NCC}, EVM_{ED}, $vt\text{EVM}_{\text{NCC}}$, $vt\text{EVM}_{\text{ED}}$, original image and PCA representations. Smaller DB-indices of clusterings on EVMs demonstrate again that the proposed representation allows for more suitable clustering of the data for all three endoscopic datasets, where a slight improvement is observed by using the S_{NCC} instead of the S_{ED}.

Examples for clustering results achieved on the EVM_{NCC} representations for three clinical datasets are presented in Figure 7.11. Figure 7.12(a), 7.12(b) and 7.12(c) demonstrate sample frames from each cluster on the EVM_{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 Figure 7.12(b) or 4th cluster in Figure 7.12(c) are observed due to some remaining uninformative frames and can be addressed by increasing the number of clusters on EVM_{EH}. However, due to the manual selection of the uninformative clusters by an

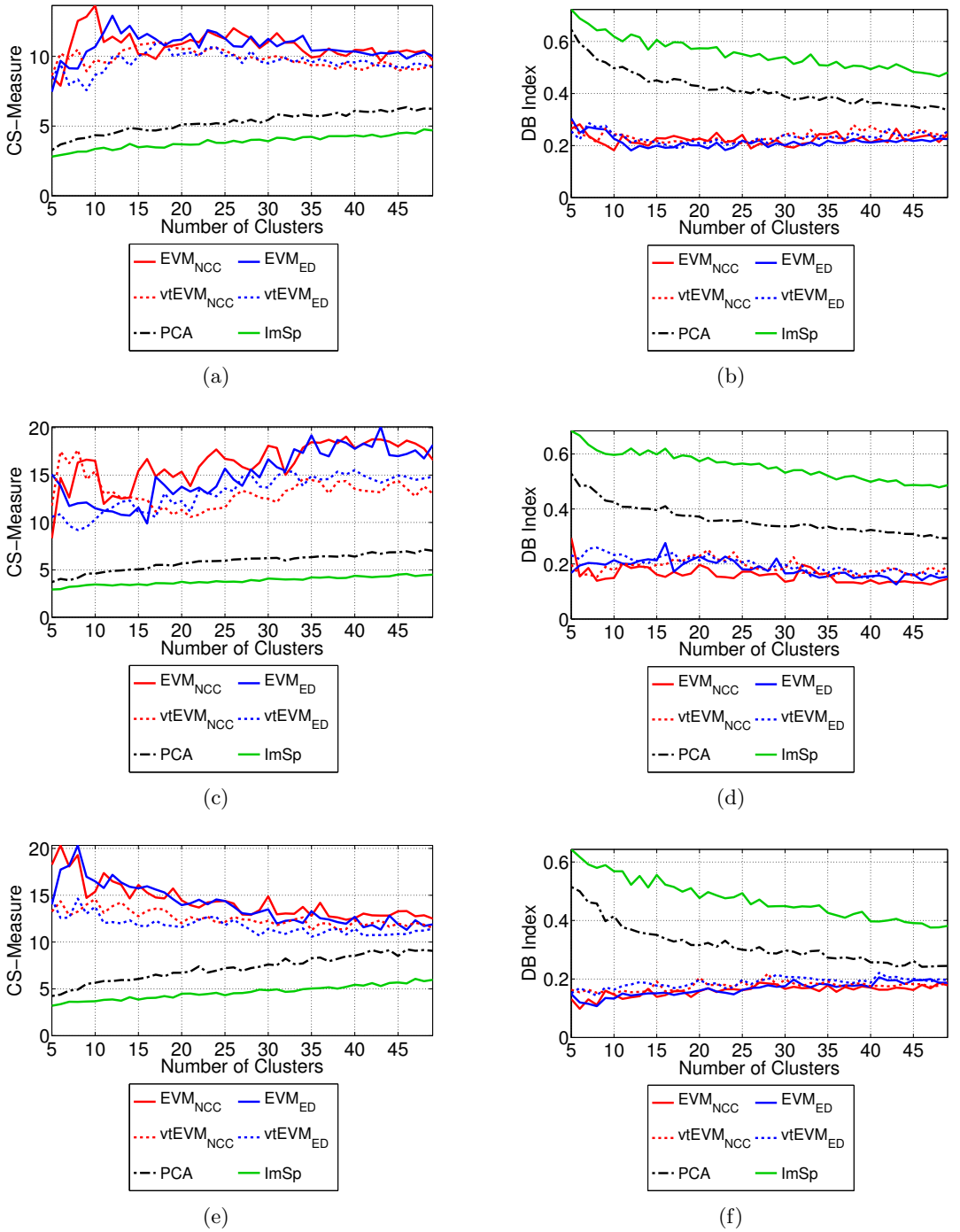
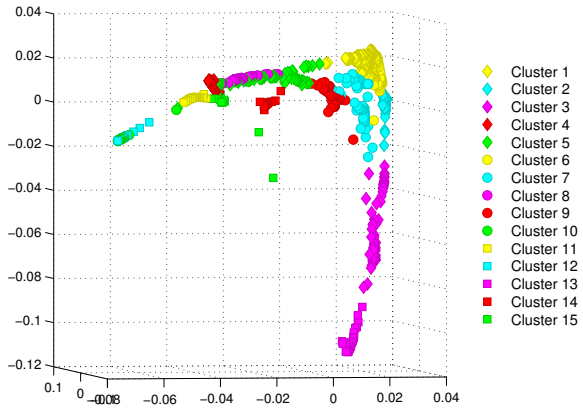
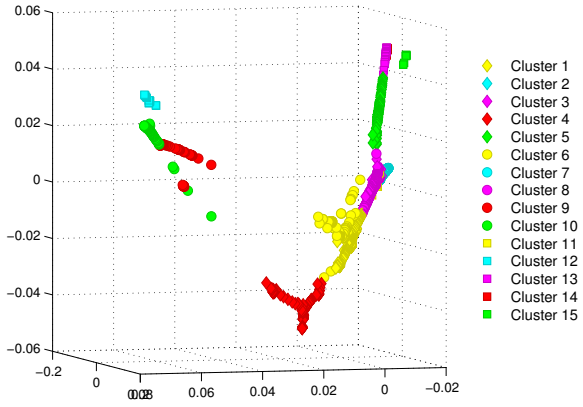


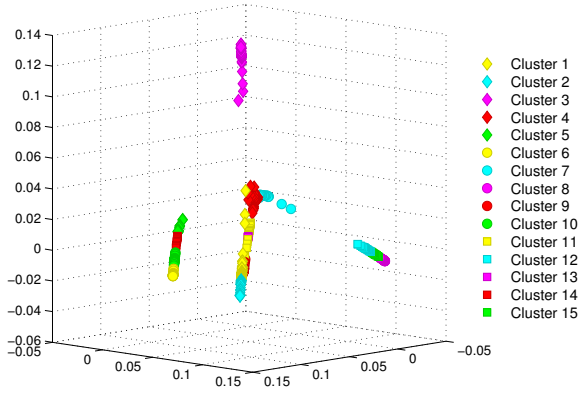
Figure 7.10: CS-measure ((a), (c) and (e)) and DB-index ((b), (d) and (f)) evaluated for K-means clustering performed in EVM_{NCC}, EVM_{ED}, $vt\text{EVM}_{\text{NCC}}$, $vt\text{EVM}_{\text{ED}}$, original image and PCA representations for different number of clusters ranging from 5 to 50 on the 1st, 2nd and 3rd endoscopic datasets.



(a)

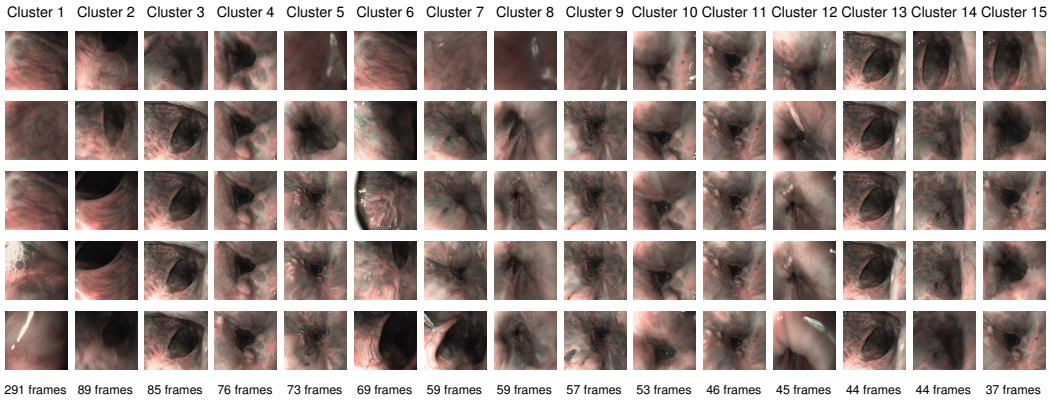


(b)

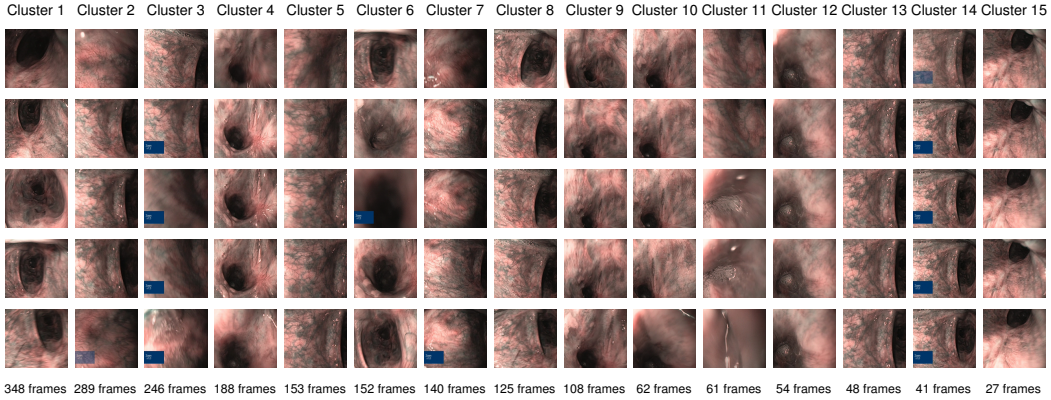


(c)

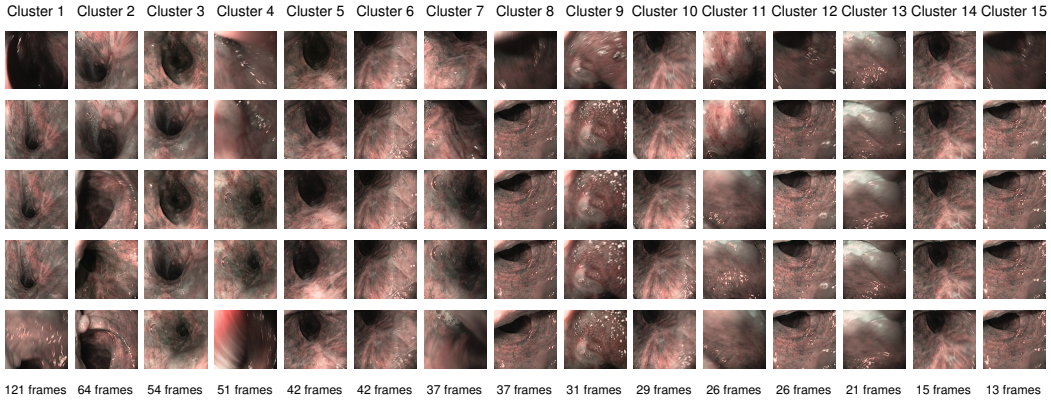
Figure 7.11: (a), (b) and (c) show the first 3-dimensions of the EVMs computed using the proposed NCC measure after eliminating the uninformative frames from the 1st, 2nd, and 3rd endoscopic video, respectively. An example clustering performed on EVM_{NCC} using K-means algorithm with 15 clusters is demonstrated as colour coding.



(a)



(b)

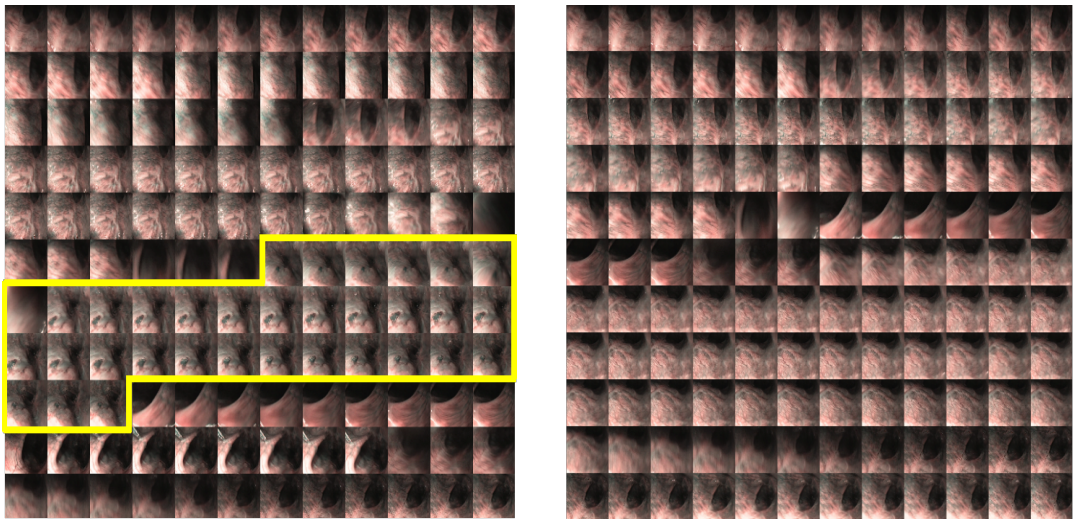


(c)

Figure 7.12: (a), (b) and (c) show the results of the clustering on EVM_{NCC} for the 1st, 2nd and 3rd dataset, respectively. For each dataset all 15 clusters are illustrated, where the 1st, 3rd and 5th rows show the first, centre and the last frames of each cluster, respectively. 2nd and 4th row show two example frames of each cluster.

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 improvement achieved by using the normalized cross correlation measure instead of the standard Euclidean distance while creating the EVM representations can be seen in Figure 7.13 . Frames of an example cluster formed on EVM_{ED} (Figure 7.13(a)) contains several outlier frames that do not belong to the same anatomical segment of the oesophagus, whereas the corresponding cluster formed on EVM_{NCC} (Figure 7.13(b)) only consists of frames showing the same anatomical region from different endoscope viewpoints.



(a) (b)

Figure 7.13: Frames of an example PSES defined by clustering performed on (a) EVMs with Euclidean distance measure (EVM_{ED}), (b) EVMs with the proposed normalized cross correlation measure (EVM_{NCC}). Outlier frames showing a different anatomical region in EVM_{ED} clusters are marked with yellow in (a).

7.4 Conclusions

In this chapter, we presented a two step clustering of the diagnostic endoscopy. The first clustering performed on EVM_{EH} created with energy histogram similarity measure $\mathcal{S}_{EH}$ leads to a clear separation of informative and uninformative frames. In several experiments, we demonstrated the effectiveness of the EVM_{EH} representation for this clustering task. The second clustering is performed only on the informative frames labelled by the endoscopic expert with the aim of defining the PSES. To this end, we introduced another EVM representation, EVM_{NCC}, created using the normalized cross

correlation measure. We quantitatively evaluated the performance of EVM representations with and without temporal constraints and provided a comparison to the original image representation and its principal components analysis. The EVM representations respect the non-linear pairwise relations between data points (frames) while mapping each data point into a low dimensional space. Thus, by construction visually different segments become better 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 our experiments.

Chapter 8

Scene Recognition in Surveillance Endoscopy

> 'The worthwhile problems are the ones you can really solve or help solve, the ones you can really contribute something to.'
>
> Richard P. Feynman

In Chapter 7, we have discussed the two step clustering of the diagnostic endoscopy in order to identify the informative frames and to define the patient specific endoscopic segments. A PSES is a collection of frames showing the same scene (anatomical part of the oesophagus) in the diagnostic endoscopy. Once the PSESs are defined by clustering the diagnostic endoscopy, classification of surveillance endoscopic frames into these PSESs allows for recognition of optical biopsy sites during the surveillance examination. In this chapter, we present two different approaches for scene recognition in surveillance endoscopy.

The first approach, presented in Section 8.1, is based on assigning each surveillance frame one PSES individually. This method does not require any additional user interaction. However, its application is limited to cases where the oesophageal tissue did not undergo significant structural changes.

In more complicated situations, such as when the patient had chemo-therapy or mucosal resection between the two examinations, the recognition (classification) problem is challenged by the severe visual changes of the oesophageal tissue, as illustrated in Figure 2.4. To facilitate the scene recognition in these highly challenging cases, we present a novel approach which relies on inter-examination cluster correspondences. Firstly, we introduce two run surveillance endoscopies for upper GI examinations, which are commonly performed and highly recommended in Bronchoscopy. Relying on

cluster correspondences between the diagnostic and the first run surveillance videos, the proposed method is able to perform scene recognition on highly challenging datasets.

This scene matching method, which extends the individual frame classification presented in Section 8.1, was first introduced in our previous study [Atasoy et al., 2011] and is explained in detail in Section 8.2 in this Chapter. Sections 8.1.3 and 8.2.4 present quantitative evaluations of the first and second classification approaches, respectively.

8.1 Classification of Individual Frames

In the literature, the focus of endoscopic image classification is mainly directed towards computer aided diagnosis of polyps [Stehle et al., 2009, Gross et al., 2009, Tamaki et al., 2010] and tumours [Karkanis et al., 2002b] or detection of endoscopic lesions [Iakovidis et al., 2006, Maroulis et al., 2003, Karkanis et al., 2002a]. Recently, video summary using representative frame extraction has also been investigated for wireless capsule endoscopy [Iakovidis et al., 2010, Iakovidis et al., 2008].

In this section, we address the identification of a new endoscopic frame with one of the PSESs via classification. In our method, 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 several frames showing the same segment of the oesophagus from different viewpoints of the endoscope as also illustrated in Figure 7.13.

8.1.1 Projection onto the EVMs

In order to perform the classification also in the low dimensional EVM representation, as used for the clustering, a new endoscopic frame has to be mapped into this low dimensional space. For non-linear manifold learning techniques, such as [Belkin and Niyogi, 2003], the mapping from the high to the low dimensional space is not readily extendible to new data points. For the Laplacian eigenmaps method [Belkin and Niyogi, 2003], a linear approximation of this mapping can be computed using the LPP method [He et al., 2005], as derived in Section 6.6.

Given the informative frames of the diagnostic endoscopy $\widehat{\mathbf{I}} = \{\widehat{\mathbf{I}}_1, \cdots, \widehat{\mathbf{I}}_p\} \subset \mathbf{I}$ and a new surveillance endoscopic frame $\mathbf{I}_s \notin \mathbf{I}$, 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 6.6. To this end, the transformation matrix $\mathbf{T}$ is computed by solving Equation (6.11) and using $\widehat{\mathbf{I}} \cup \mathbf{I}_s$ as the input dataset. This allows us to define a mapping function from the high to the low dimensional space and to optimize the linear mapping function only for the informative frames of the diagnostic endoscopy together with the new frame coming from the surveillance endoscopy. Then, we project each data point $\mathbf{I}_i$ in the training dataset $\mathbf{I}$ into the low dimensional space

using this linear transformation as:

$$\mathbf{y}_i = \nu(\mathbf{I}_i) = \mathbf{T}^\top \mathbf{I}_i \ . \tag{8.1}$$

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

8.1.2 Assigning a PSES to a Query Frame

In the final classification, we classify a new frame $\mathbf{I}_s \notin \mathbf{I}$ as informative/uninformative and assign it to a PSES (if informative). To do so, we consider as classes the eliminated uninformative clusters (as explained in Section 7.2)

$$\overline{\Omega^{\mathrm{EH}}} = \{\overline{C_1}^{\mathrm{EH}}, \cdots, \overline{C_\alpha}^{\mathrm{EH}}\} \ , \tag{8.2}$$

together with the PSESs (as explained in Section 7.3)

$$\Omega^{\mathrm{PSES}} = \{C_1^{\mathrm{PSES}}, \cdots, C_\beta^{\mathrm{PSES}}\} \ . \tag{8.3}$$

Thus, the final set of classes to assign is formed as:

$$\Omega = \{\overline{\Omega^{\mathrm{EH}}}, \Omega^{\mathrm{PSES}}\} \ . \tag{8.4}$$

To classify a new frame $\mathbf{I}_s$, its NN is found among the frames of the diagnostic endoscopy in the low dimensional representation $\nu(\mathbf{I}_s)$. The class of the NN($\mathbf{I}_s$) is also assigned to the surveillance endoscopic frame $\mathbf{I}_s$. 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 PSES of the diagnostic endoscopy. If a PSES containing an optical biopsy location is assigned to a new surveillance endoscopy frame, the endoscopist can be notified online during the examination. The proposed framework consisting of clustering of the diagnostic endoscopy end classification of the individual surveillance endoscopy frames is illustrated in Figure 8.1.

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 [Mountney et al., 2009, Allain et al., 2010, Allain et al., 2009, Atasoy et al., 2009].

8.1.3 Evaluation

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.

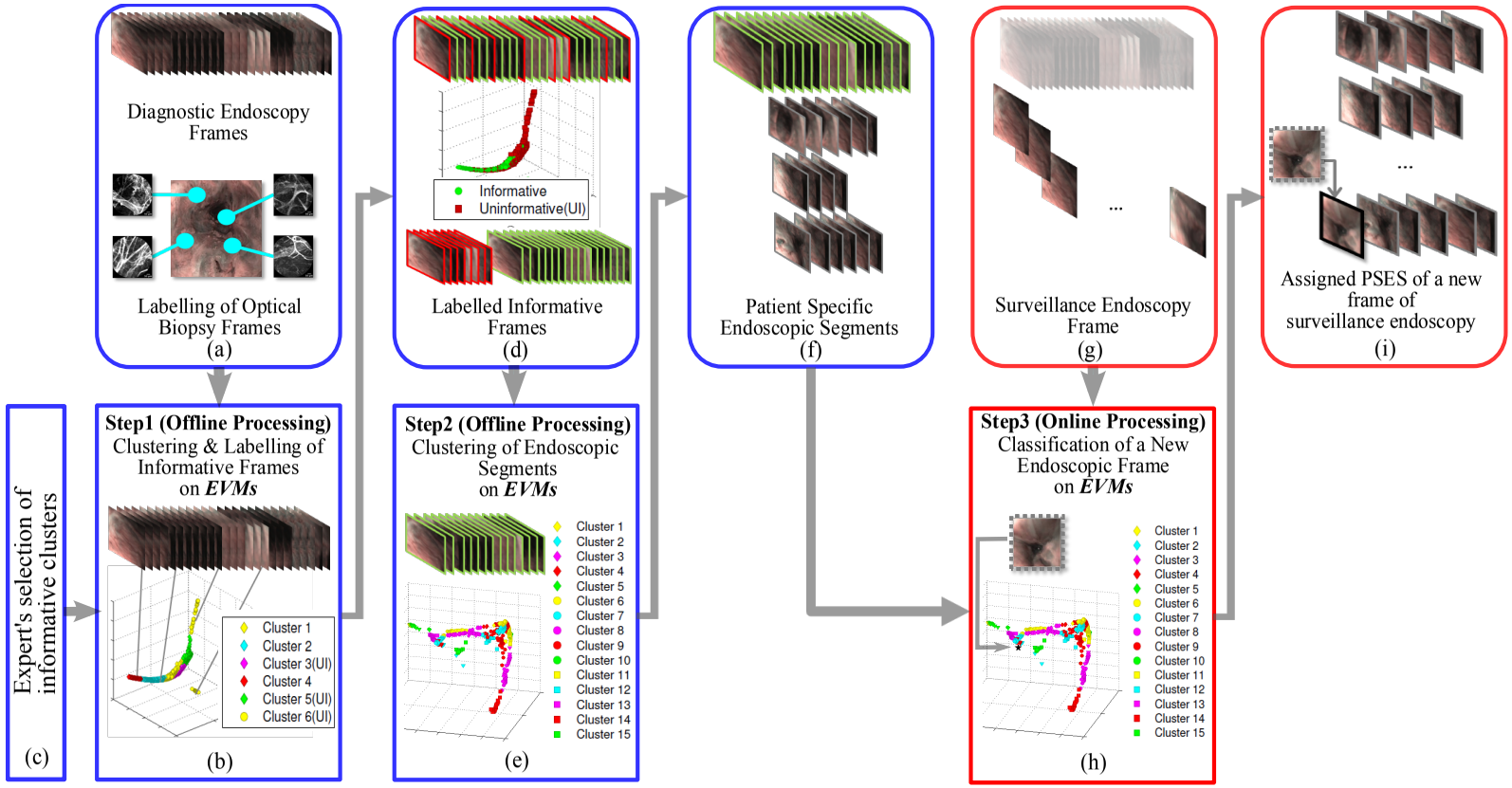


Figure 8.1: Framework for classification of individual frames. (a)-(f) represent steps of the offline clustering part as explained in Chapter 7 and illustrtaed in Figure 7.2. (f) PSESs corresponding to clusters of informative frames. Together with previously labelled uninformative clusters, PSESs form 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 neighbour classifier in the low dimensional space. (i) Outcome of the last step of the framework; assigned PSES of the new endoscopic frame.

After defining the PSESs by clustering endoscopic frames on EVMs, classification of a new frame into one of these defined PSESs is performed by a simple NN classifier as explained in Sections 8.1.1 and 8.1.2. The accuracy of the classification step largely depends on the definition of the PSESs and thus on the clustering of the diagnostic dataset.

8.1.3.1 Evaluation on Informative Frames

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 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 before classifying each frame of the endoscopic datasets.

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 using the PSESs defined on six representations, as listed in Section 7.3.3.2, are again compared. Figures 8.2(a), 8.2(c) and 8.2(e) show the classification accuracy for all representations for different number of PSESs ranging from 5 to 50. Mean and standard deviation for classification accuracies over all number of clusters is shown in Figures 8.2(b), 8.2(d) and 8.2(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, EVM_{NCC} results in slightly more accurate classification as compared to the EVM_{ED}.

8.1.3.2 Evaluation on Complete Patient Datasets

In this final experiment, we perform a quantitative evaluation of the entire individual frame classification 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 $\{\overline{C_1}^{\text{EH}}, \cdots, \overline{C_\alpha}^{\text{EH}}\}$ and the *PSESs* $\{C_1^{\text{PSES}}, \cdots, C_\beta^{\text{PSES}}\}$ are merged as explained in Section 8.1.2. Like in the classification experiments on informative frames, we perform a LOPO evaluation on the complete endoscopic dataset, this time also including the labelled uninformative frames into the

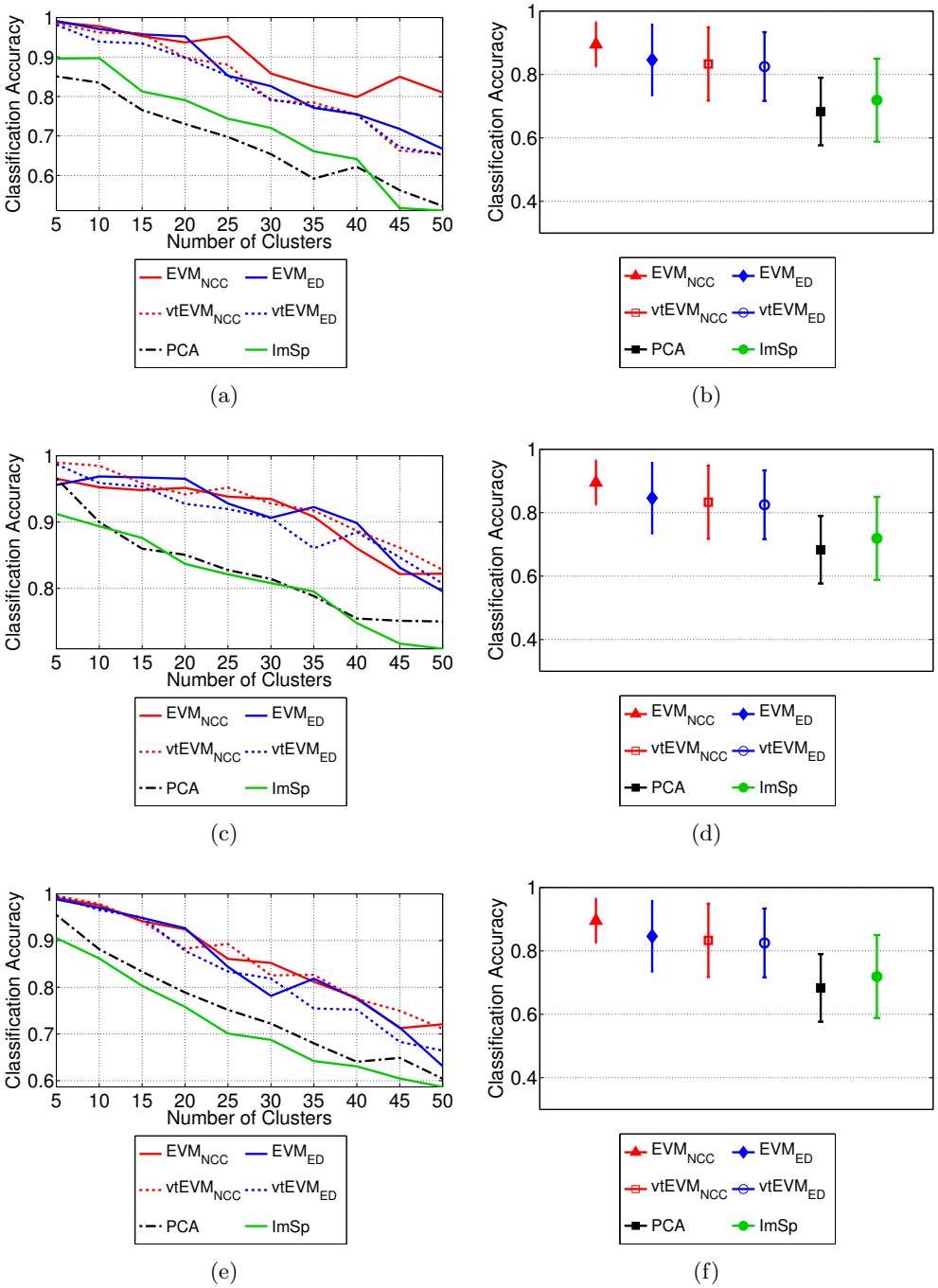


Figure 8.2: Classification accuracy estimated by LOPO experiments on the (a) 1st, (c) 2nd and (e) 3rd dataset using the PSESs defined on the informative frames only. The experiments are carried out for different number of PSESs ranging from 5 to 50 defined as the clusterings in image space, PCA, EVM_{ED}, EVM_{NCC}, vtEVM_{ED} and $\text{vtEVM}_{\text{NCC}}$. (b), (d) and (f) show the mean and standard deviation of classification accuracies over all number of PSESs for the 1st, 2nd and 3rd dataset, respectively.

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 the classification accuracy with number of PSES ranging from 5 to 50 for all three datasets is presented in Figures 8.3(a), 8.3(c) and 8.3(e). The mean and standard deviation of the classification accuracies over the number of PSES is shown in Figures 8.3(b), 8.3(d) and 8.3(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.

8.2 Classification with Scene Correspondences

The individual frame classification approach introduced in Section 8.1 provides the first step towards a scene recognition framework for re-targeting optical biopsy sites in surveillance examinations. Its online application during the surveillance procedure does not require any user interaction. However, it can not account for significant changes of the tissue appearance between the diagnostic and surveillance examinations. Therefore, this frame classification approach is suitable for endoscopic examinations of patients with early stage diagnosis such as BO.

In this section, we extend the individual frame classification by an inter-examination scene matching. To this end, a pre-examination endoscopic video is acquired before performing the actual surveillance examination. Like the diagnostic endoscopy ($\mathcal{D}$), this first run surveillance endoscopy ($\mathcal{S}1$) is clustered into different PSESs and correspondences between the PSESs of $\mathcal{D}$ and $\mathcal{S}1$ are selected by the endoscopic expert. During the actual examination, which is performed in the second run surveillance endoscopy ($\mathcal{S}2$), these correspondences serve as the bridge between the $\mathcal{S}2$ frames and PSESs of the diagnostic endoscopy. A new endoscopic frame is first assigned a PSES from the $\mathcal{S}1$ and then, relying on the cluster correspondences between $\mathcal{D}$ and $\mathcal{S}1$, the corresponding PSES of the diagnostic endoscopy is retrieved.

Next, we explain the individual steps of the proposed framework using scene correspondences.

8.2.1 Two-run Surveillance Endoscopy

Currently available treatment techniques for the oesophageal cancer such as chemotherapy or mucosal resection can lead to significant changes in the structure of the oesophageal tissue. The major challenge in performing scene recognition between the diagnostic and surveillance endoscopies of these patients is the variation in visual appearances of the same scene as demonstrated in Figure 2.4. To address this challenge, we propose a *two-run surveillance endoscopy*. In the introduced workflow, prior to the actual surveillance endoscopy, a first-run surveillance ($\mathcal{S}1$) video is acquired in the same examination. This is a commonly performed and highly recommended process in bronchoscopy [Häussinger et al., 2004]. For GI examinations, however, this process has not yet been utilized. For extending the individual frame classification, we introduce

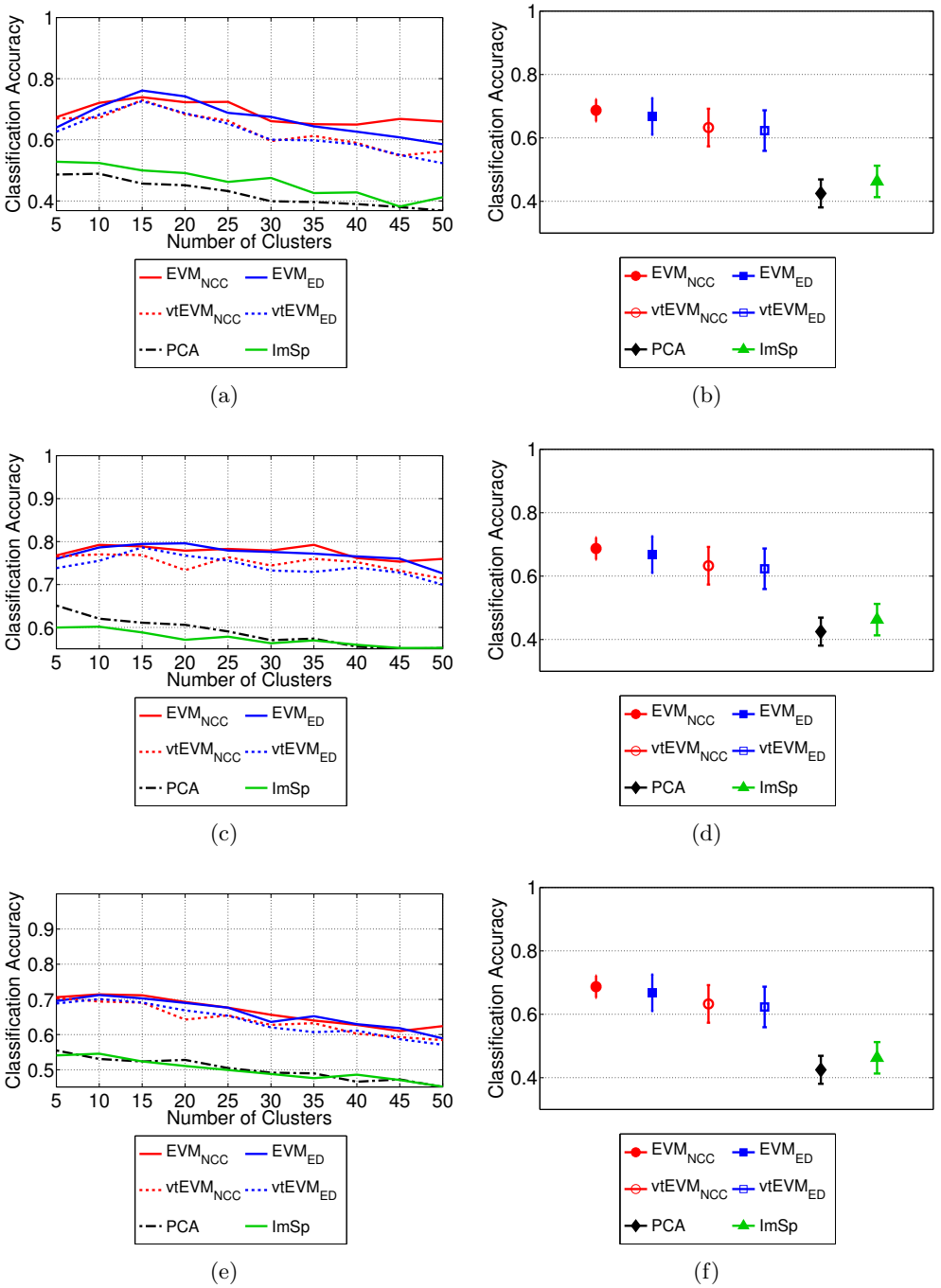


Figure 8.3: Classification accuracy estimated by LOPO experiments on the (a) 1st, (c) 2nd and (e) 3rd dataset including the uninformative clusters. The experiments are carried out for different number of PSESs ranging from 5 to 50 defined as the clusterings in image space, PCA, EVM_{ED}, EVM_{NCC}, vtEVM_{ED} and $\text{vtEVM}_{\text{NCC}}$. (b), (d) and (f) show the mean and standard deviation of classification accuracies over all number of PSESs for the 1st, 2nd and 3rd dataset, respectively.

the two run surveillance schema for GI endoscopies. This allows us to provide an applicable solution for re-targeting the optical biopsy sites in challenging surveillance examinations.

In the introduced two run surveillance endoscopy schema, prior to the actual surveillance endoscopy, the endoscope is guided from the mouth to the z-line (junction from the oesophagus to the stomach) without acquiring any optical biopsies. The video of this $\mathcal{S}1$ endoscopy is clustered into different endoscopic scenes and used to acquire scene matching between the diagnostic and surveillance endoscopy. This additional step enables the recognition of the same location despite very large variation in the visual appearances of the scene in different examinations, as illustrated in Figures 8.4(a), 8.4(g).

The proposed workflow for this extended classification method with scene correspondences involves 3 endoscopic videos:

- diagnostic ($\mathcal{D}$) endoscopy, where the first optical biopsies have been acquired (Figure 8.4(a));
- first run surveillance ($\mathcal{S}1$) endoscopy, which is performed to provide scene matches between two different endoscopic examinations (Figure 8.4(d)); and
- second run surveillance ($\mathcal{S}2$) endoscopy, where the surveillance examination is performed and the previous optical biopsy sites need to be recognized in real-time and *in-vivo* (Figure 8.4(g)).

The proposed workflow consists of the following main steps:

1. Clustering of the $\mathcal{D}$ endoscopy into PSESs (Figure 8.4(a)-8.4(c)),
2. Acquisition of the $\mathcal{S}1$ endoscopy (Figure 8.4(d)),
3. Clustering of the $\mathcal{S}1$ endoscopy into PSESs (Figure 8.4(d)-8.4(f)),
4. Selection of the query (optical biopsy) clusters in $\mathcal{D}$ endoscopy and their correspondences in the $\mathcal{S}1$ by the endoscopic expert,
5. Nearest neighbour matching and $\mathcal{S}1$ cluster assignment to each frame of the $\mathcal{S}2$ endoscopy in real-time (Figure 8.4(g)),
6. Notification of the expert during the $\mathcal{S}2$ endoscopy if a frame is assigned to one of the (query) optical biopsy clusters.

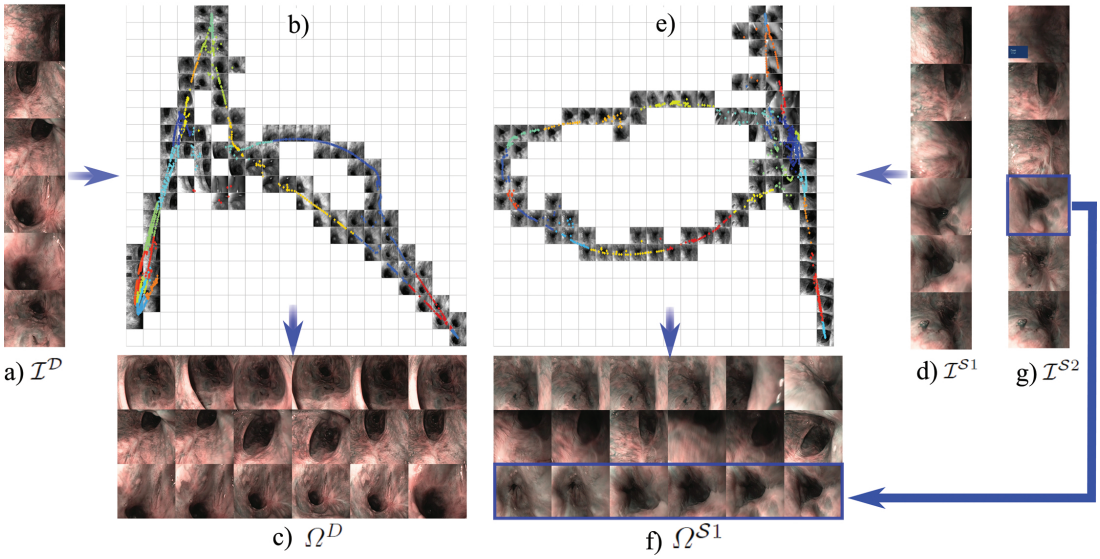


Figure 8.4: Proposed workflow. (a) Example frames from the diagnosis endoscopy. (b) the 1st and 2nd dimensions of the manifold of the diagnostic endoscopy created using the visual and temporal relations within the video. Frames showing similar locations are clustered together, where different clusters are illustrated with different colours. (c) Example clusters of the diagnostic endoscopy, where different columns correspond to different clusters. Note that frames of the same scene with different endoscope viewpoint are clustered together whereas different scenes are clustered separately. (d) Corresponding scenes of (a) in the $\mathcal{S}1$. Rows in (a) correspond to rows in (d). (e) 1st and 2nd dimensions of $\mathcal{S}1$ manifold and the computed clusters. (f) Example frames of the corresponding clusters of (c) in the $\mathcal{S}1$. The columns in (c) correspond to columns in (f). (g) Example frames used as the $\mathcal{S}2$ endoscopy in our experiments.

8.2.2 Scene Clusters

Given the frames of the $\mathcal{D}$ endoscopy (Figure 8.4(a)) and of the $\mathcal{S}1$ endoscopy (Figure 8.4(d)), we first compute the EVM representation for each video by taking into account the visual similarities and the temporal relations between the frames as discussed in Section 6.3. Figure 8.4(b) and 8.4(e) show the 1st and 2nd dimensions of the manifolds computed from $\mathcal{D}$ and $\mathcal{S}1$ endoscopies respectively, where the clusters are illustrated by different colours. Clustering of the frames into PSESs is performed on this manifold representation.

In the experiments carried out in Section 8.2.4, we use the finite mixture models (FMM) method proposed in [Figueiredo and Jain, 2002] to compute the clusters. This clustering approach is based on a mixture model and the expectation maximization and estimates the probability $P[c(\mathbf{y_i}) = C_j]$ of each point $\mathbf{y_i}$ belonging to a mixture model (cluster) C_j. Once the probabilities for each data point $\mathbf{y_i}$ and each cluster C_j

are computed, we assign the cluster with the highest probability

$$c(\mathbf{y_i}) = \underset{C_j}{\operatorname{argmax}}\, P[c(\mathbf{y_i}) = C_j)] \ . \tag{8.5}$$

This method can be seen as an extension of the K-means algorithm [Hartigan and Wong, 1979] to anisotropic distributions of the data points. The K-means clustering algorithm relies on the assumption that the data points are distributed according to an isotropic Gaussian distribution, whereas the FMM method is also suitable for anisotropic distributions. Qualitatively, we did not observe significant differences in the results of the K-means and FMM clusterings, which can indicate that the isotropic distribution as assumed by the K-means algorithm is sufficient for modelling the data distribution on the EVMs.

Figure 8.4(c) shows example clusters from $\mathcal{D}$ endoscopy where the corresponding clusters in the $\mathcal{S}1$ are illustrated in Figure 8.4(f). Each row illustrates a different cluster, where corresponding rows in Figures 8.4(c) and 8.4(f) demonstrate corresponding scenes (clusters) in $\mathcal{D}$ and $\mathcal{S}1$. Note the severe change in the appearance of the scenes between the two examinations. Based on the previously defined $\mathcal{D}$ endoscopy clusters and their correspondences in the $\mathcal{S}1$, the proposed workflow allows for *real-time* and *in-vivo* recognition of the query scenes during the $\mathcal{S}2$.

8.2.3 Scene Recognition

After defining the PSESs of the diagnostic endoscopy

$$\Omega^{\mathcal{D}} = \{C_1^{\mathcal{D}}, \cdots, C_\alpha^{\mathcal{D}}\} \ , \tag{8.6}$$

and then the ones of the $\mathcal{S}1$ endoscopy

$$\Omega^{\mathcal{S}1} = \{C_1^{\mathcal{S}1}, \cdots, C_\beta^{\mathcal{S}1}\} \ , \tag{8.7}$$

both clusterings are provided to the endoscopic expert. The set of Q clusters, where an automatic recognition is needed, *i.e.* the query (optical biopsy) clusters

$$\{C_q^{\mathcal{D}}\}_{q=1}^{Q} \in \Omega^{\mathcal{D}} \ , \tag{8.8}$$

as well as their correspondences in the $\mathcal{S}1$ endoscopy,

$$\{C_{\gamma(q)}^{\mathcal{S}1}\} \in \Omega^{\mathcal{S}1} \ , \tag{8.9}$$

(where γ denotes the correspondence relation) are selected by the endoscopic expert.

During the $\mathcal{S}2$, first the image closest to a frame $\mathcal{I}_i^{\mathcal{S}2}$, that is $\mathcal{I}_j^{\mathcal{S}1} = \mathrm{NN}(\mathcal{I}_i^{\mathcal{S}2})$, is found by a simple NN matching using Euclidean distances. Then each frame $\mathcal{I}_i^{\mathcal{S}2}$ is assigned the cluster of its NN

$$c^{\mathcal{S}1}(\mathcal{I}_i^{\mathcal{S}2}) = c^{\mathcal{S}1}(\mathcal{I}_j^{\mathcal{S}1}) \ , \tag{8.10}$$

and, by transition, the corresponding diagnosis endoscopy cluster

$$c^{\mathcal{D}}(\mathcal{I}_i^{\mathcal{S}2}) = c^{\mathcal{D}}(\mathcal{I}_j^{\mathcal{S}1}) \ . \tag{8.11}$$

If a frame is determined to belong to a query cluster

$$c^{\mathcal{D}}(\mathcal{I}_i^{\mathcal{S}2}) \in \{C_q^{\mathcal{D}}\} \ , \tag{8.12}$$

the expert is notified and all frames of the corresponding diagnostic endoscopy cluster

$$\{\mathcal{I}_k^{\mathcal{D}} | c^{\mathcal{D}}(\mathcal{I}_k^{\mathcal{D}})\} \tag{8.13}$$

are retrieved.

This proposed workflow allows for including the expert's supervision in defining the query scenes and their correspondences in the $\mathcal{S}1$ without involving any training. This is an important property, since long training processes, such as performed in several learning-based methods, would not be feasible for routine clinical applications.

8.2.4 Evaluation

For quantitative evaluation of the proposed classification approach with scene correspondences, we perform 3 experiments. In each experiment, 40 frames from the surveillance endoscopic video are selected by regularly sampling the frames over time and are used as test frames simulating the $\mathcal{S}2$ endoscopy. This leads to a total recognition of 120 scenes in the experiments. Remaining parts of the surveillance video are defined to be the $\mathcal{S}1$ endoscopy.

For these experiments, we use vtEVM$_{\text{ED}}$ as the representative EVM representation where the clustering is performed in the low dimensional EVM representation estimated using the optimal linear mapping, as discussed in Section 6.6 and in [Atasoy et al., 2011, He et al., 2005]. We compare the presented classification with scene correspondences to the direct classification of $\mathcal{S}2$ frames in the original image representation. To this end, we firstly classify a $\mathcal{S}2$ endoscopic frame into the PSESs defined on $\mathcal{S}1$ endoscopy. Then, using the inter-examination scene correspondences between the PSESs of $\mathcal{S}1$ and $\mathcal{D}$, we retrieve all frames of the corresponding cluster in $\mathcal{D}$ endoscopy. For comparison, we retrieve m-NN of a test frame $\mathcal{I}_i^{\mathcal{S}2}$ among the $\mathcal{D}$ frames directly. For a fair evaluation, we choose m to be equal to the number of frames retrieved with scene correspondences and perform the NN classification for both methods using the Euclidean Distances in the image space.

For the evaluation, true positives (tp) and false positives (fp) are determined by expert visual inspection of the retrieved frames. The false negatives (fn) of each method are defined relatively, as the number of frames that one method is able to correctly retrieve but not the other. Recall ($tp/(tp+fn)$) and precision ($tp/(tp+fp)$) values are evaluated for each test frame and mean and standard deviation achieved by both methods are presented in Figure 8.5.

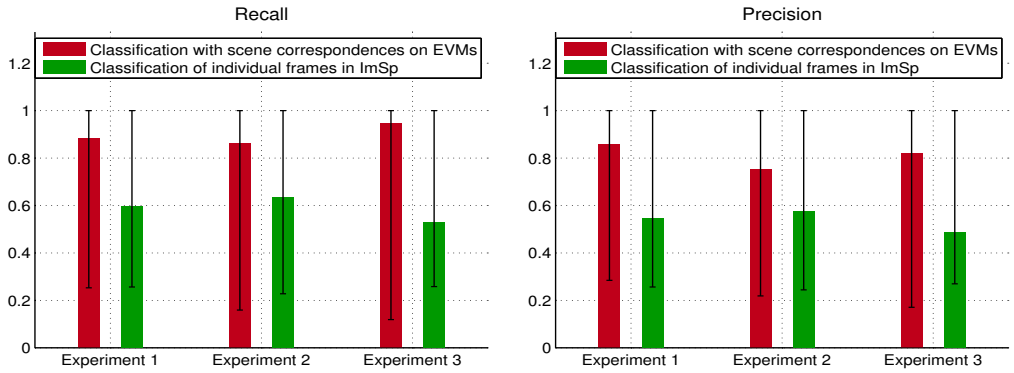


Figure 8.5: Mean and standard deviation of recall and precision of the proposed method and of the direct application of the m-NN matching to the diagnosis endoscopy are evaluated relative to each other by visual assessment of the retrieved frames.

Application of the m-NN matching in the original image representation directly between the test frames and the diagnostic endoscopy results in only 58.54% mean recall and 53.58% mean precision. The presented classification method leads to a 89.75% recall and 80.91% precision on average using the same NN matching between the test frames and the $\mathcal{S}1$ endoscopic frames and then applying the cluster correspondences. Examples of the correctly recognized frames using the proposed method in comparison to the direct application of m-NN matching between the $\mathcal{S}2$ and $\mathcal{D}$ videos are demonstrated in Figure 8.6. Due to the use of the EVM representation, the formed endoscopic clusters contain frames showing the same location from different viewpoints and from different parts of the video. This leads to creation of meaningful clusters in the $\mathcal{D}$ and $\mathcal{S}1$ endoscopies and thus results in more accurate classification as reflected in the high recall and precision values of the proposed method in Figure 8.5.

8.3 Conclusions

In this Chapter, we first presented an approach for classification of individual frames (Section 8.1) and also investigated the effect of including temporal constraints into the EVM framework. All proposed EVM representations; i.e. EVM_{NCC}, EVM_{ED}, $vt\text{EVM}_{\text{NCC}}$ and $vt\text{EVM}_{\text{ED}}$ show comparable accuracy in classification experiments. Quantitative evaluation of the individual frame classification is performed in two different LOPO experiments. The first experiment is carried out using the informative frames labelled as explained in Section 7.2, whereas the second experiment included both informative and uninformative frames in the diagnostic as well as surveillance datasets. Both experiments demonstrated that the proposed EVM representation yields higher accuracy 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 on the EVMs. In addition, having much lower dimensionality than

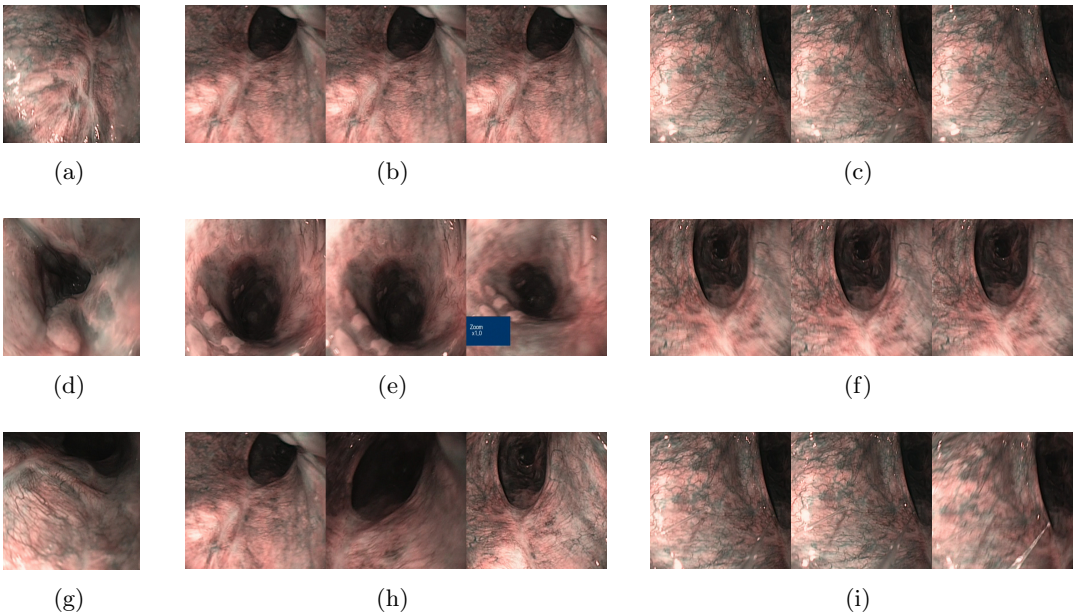


Figure 8.6: (a), (d) and (g) show test frames used as $\mathcal{S}2$ endoscopy. (b), (e) and (h) illustrate correctly recognized frames using the presented classification approach with scene correspondences and two run surveillance endoscopies for (a), (d) and (g) respectively. (c), (f) and (i) illustrate the first three NN retrieved by direct matching to the diagnostic endoscopy for (a), (d) and (g) respectively. Note that the presented approach is able to retrieve the correct frames whereas the direct k-NN matching in the original image representation fails due to the significant change in endoscope viewpoint in (a) and (g) and the structural changes in (d).

the original image representation, EVMs provide a means to perform clustering and classification tasks in a more efficient manner.

In Section 8.2, we presented an endoscopic scene recognition method based on two run surveillance endoscopies and correspondences of scene clusters (PSESs). The proposed method enables the re-targeting of the optical biopsy locations in *surveillance* endoscopies despite significant structural changes of the oesophageal tissue. The introduced workflow with two run surveillance endoscopies creates a link between the scenes of the diagnostic and surveillance examinations. This reformulation reduces the very challenging inter-examination re-targeting into the plausible problem of intra-examination frame recognition. The performance of the presented classification method is quantitatively evaluated on three different datasets and compared to a direct m-NN matching between the $\mathcal{D}$ and $\mathcal{S}2$ videos. The superior results achieved by the presented approach demonstrate its feasibility to recognize the optical biopsy scenes even in challenging surveillance datasets and provide an encouraging step towards its application in the daily clinical routine.

Chapter 9

Intra-Frame Localisation

> 'Space is not a lot of points close together; it is a lot of distances interlocked.'
>
> Sir Arthur Stanley Eddington

This chapter presents an intra-frame localisation approach based on deformable wide baseline matching. Once an optical biopsy frame is recognized within the surveillance endoscopic video, this method provides a way for establishing point correspondences between the two endoscopic frames. To this end, firstly distinctive local image regions are detected in both frames individually. These regions are described using scale invariant feature transform (SIFT). After computing the viewpoint invariant descriptor vectors, the matching problem is formulated as a Markov random field (MRF) model. Thereby, a new pairwise constraint is proposed, as explored previously in our study [Atasoy et al., 2009], which is invariant to large degree of tissue deformation and endoscope view point change.

In Sections 9.1 and 9.2 in this Chapter, we explain the individual steps for our deformable wide-baseline matching method. Quantitative and qualitative evaluation of the introduced method and final conclusions are presented in Section 9.3 and 9.4, respectively.

9.1 Affine Covariant Region Detection and Description

In the recent literature, detection of distinctive local image features and their description in a manner that is invariant to a class of transformations has received significant attention in the computer vision community (For an overview of affine covariant region detection and description methods, we refer to [Mikolajczyk et al., 2005] and

[Mikolajczyk and Schmid, 2005] respectively). The main motivation behind the development of these techniques was the search for an image representation which allows for recognition of objects or scenes despite large variations in the imaging conditions such as illumination or camera viewpoint change.

In order to address the correspondence problem between two images which differ by a large perspective transformation, affine covariant features have been proposed and successfully applied for several tasks including scene and object recognition [Lowe, 2004, Lowe, 1999, Sivic and Zisserman, 2009] and microscopic image retrieval [André et al., 2009a]. Covariant regions with a class of transformations can be defined as two local regions corresponding to the same 3D surface patch (where the two input images differ by a transformation of this class). Generally, such surface patches are extracted by first detecting distinctive image regions (feature points) on a given image and then assigning to each local image feature an elliptical region depending on the viewpoint conditions.

In our intra-frame localisation method, we detect affine covariant regions independently on both images using affine invariant anisotropic region detector [Giannarou et al., 2009]. This feature detector has been shown to be robust against small deformations. For viewpoint invariant description, each elliptical region p is normalized by the corresponding affine transformation $\mathbf{M}_p$ (determined by the shape of the ellipse) and mapped onto the corresponding circular region $\overline{p} = \mathbf{M}_p p$ (Figure 9.1). Then, the dominant gradient orientation θ_p is estimated from the local image gradients and the SIFT descriptor [Lowe, 2004] $\mathbf{d}(\overline{p}, \overline{\sigma_p}, \overline{\theta_p})$ is computed from the circular patch $\overline{p}$ using the characteristic scale $\sigma_{\overline{p}}$ and the dominant gradient orientation $\theta_{\overline{p}}$. This results in a 128 dimensional descriptor vector to represent each patch in an invariant manner to affine transformations.

For computer vision applications such as scene recognition or object detection, several methods rely on the discriminative power of these region descriptors. By not accounting for the location information or the geometric relations between the feature points, the image is represented as a "bag-of-features", which yields the required invariance to changes of the camera viewpoint.

However, the lack of geometric information necessitates the detection of distinctive local features, which the GI endoscopic images lack in general. The scenes encountered in a typical GI endoscopy video consists of repetitive vascular patterns and homogeneous regions.

In order to tackle the region matching problem on such challenging endoscopic frames, we introduce an MRF model that combines the discriminative power of the local feature descriptors with the local geometry of the detected features. To this end, we propose a new pairwise cost which evaluates the geometry within a local neighbourhood based on the photometric properties of the extracted patches.

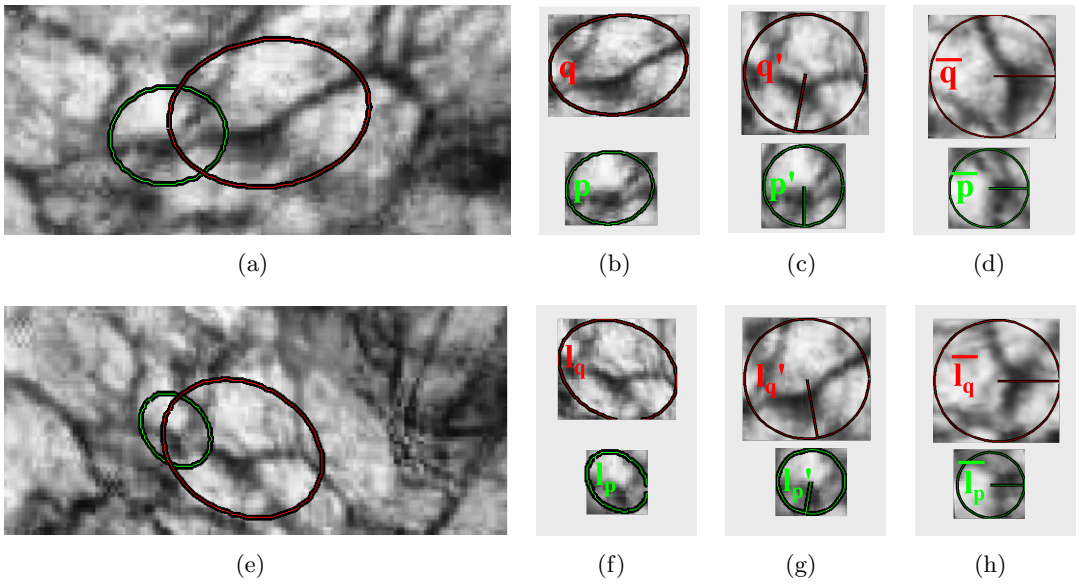


Figure 9.1: Viewpoint invariant region description. (a) and (e) show two examples for the detected regions by applying the affine co-variant region detector on the first and second input images, respectively. (b) and (f) demonstrate the detected regions before normalization. (c) and (g) illustrate mapping of the elliptical regions onto the circular domain. The determined dominant gradient orientation is illustrated with the coloured bars within the regions. (d) and (h) show the normalized regions on which the SIFT detectors are computed. Corresponding regions are illustrated with the same colour.

9.2 Markov Random Field Model

MRFs provide a powerful tool for several image analysis problems, especially when the contextual relations between some defined image features have to be taken into account. An image processing task can be posed within the MRF framework by defining the nodes and the relations between them on a graph structure, and the labels to assign to each node. The solution is given by the optimum labelling of the created MRF model; i.e. by the assignment of a label to each node while minimizing the global energy of the MRF model. For an in depth study of MRFs in image processing, we refer to [Li, 2009].

Given the computed region descriptors, we model the matching problem as global optimization of an MRF labelling. We define the regions in the first image to be the nodes $\mathscr{G} = \{1, \cdots, n\}$ of the MRF and the regions in the second image to be the labels $\mathscr{L}^+ = \{1, \cdots, m\}$ including the null-label l_0, which is assigned to regions without true correspondence in the second image. We consider up to pairwise relations. Thus, finding the maximum a posteriori (MAP) estimate of the optimum labelling l^* is equivalent to minimizing the following energy (objective) function:

$$\mathcal{E}_{\text{MRF}} = \sum_{p \in \mathscr{G}} V_p(l_p) + \sum_{p \in \mathscr{G}} \sum_{q \in \mathcal{N}_{\text{MRF}}(p)} V_{pq}(l_p, l_q) \ , \tag{9.1}$$

where $V_p(l_p)$ is the unary cost of assigning the label l_p to the node p, $V_{pq}(l_p, l_q)$ is the pairwise cost of jointly assigning the labels l_p and l_q to the nodes p and q, and $\mathcal{N}_{\text{MRF}}$ defines the neighbourhood system[1].

9.2.1 Unary Costs

In our model, photometric similarities between the node and the label regions are evaluated via the unary costs. To this end, the cost $V_p(l_p)$ is defined to be the distance of the SIFT descriptors of the node $\overline{p}$ and label $\overline{l_p}$ regions (Figure 9.2(a)).

We further define the cost $V_p(l_0)$ of assigning the null-label l_0 to a node $\overline{p}$ to be a function of the photometric similarities. The motivation is that assigning the null-label l_0 to a region that has a strong correspondence in the second image should have a higher cost than assigning it to a region with no (strong) correspondence. We define the null-cost function of the node $\overline{p}$ as:

$$V_p(l_0) = \alpha(1 - \min(V_p(\cdot))) \ , \tag{9.2}$$

where $\min(V_p(\cdot))$ is the minimum cost of assigning a label to the node $\overline{p}$, and α is the factor regulating the trade-off between the quality and the number of matches. In our experiments, the best performance is achieved for $\alpha = 0.5$ for all our in-vivo datasets. The final unary costs are computed as:

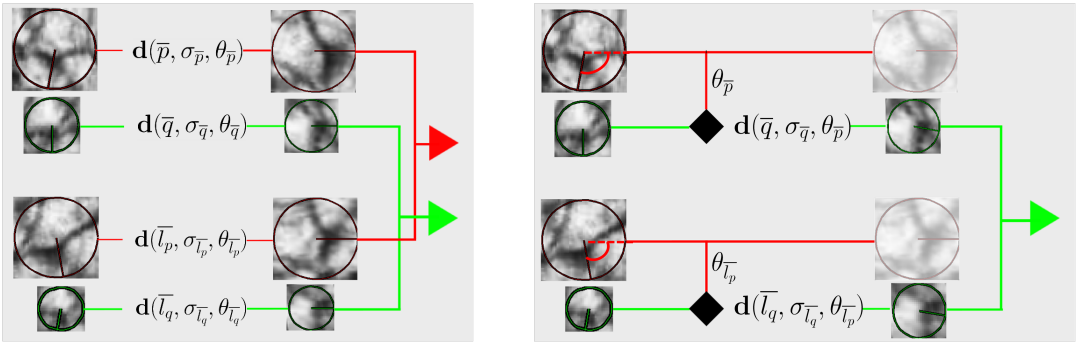
$$V_p(l_p) = \begin{cases} \angle\left(\mathbf{d}(\overline{p}, \sigma_{\overline{p}}, \theta_{\overline{p}}), \mathbf{d}(\overline{l_p}, \sigma_{\overline{l_p}}, \theta_{\overline{l_p}})\right) & \text{if } l_p \neq l_0 \\ \alpha(1 - \min(V_p(\cdot))) & \text{otherwise} \ , \end{cases} \tag{9.3}$$

where $\angle(\cdot, \cdot)$ denotes the angle between the descriptor vectors computed as $\arccos\left(\langle \mathbf{d}(\overline{p}, \sigma_{\overline{p}}, \theta_{\overline{p}}), \mathbf{d}(\overline{q}, \sigma_{\overline{q}}, \theta_{\overline{q}})\rangle\right)$. All costs $V_p(\cdot)$ are normalized to the interval $[0, 1]$ by dividing by the maximum possible angle between the two descriptor vectors.

9.2.2 Neighbourhood Systems

In the context of the matching problem, each region is allowed to have at most one correspondence in the second image, i.e. each label can be assigned at most to one node. This *uniqueness constraint* is included into the energy function by connecting each node with all the other nodes within the *global neighbourhood system* $\mathcal{N}_{\text{MRF}}$ and by defining the pairwise cost for assigning the same label to two different nodes to be infinite:

[1] It is important to note that the neighbourhood system $\mathcal{N}_{\text{MRF}}$ refers to the *neighbourhood within one endoscopic frame* and differs from the neighbourhood $\mathcal{N}$ used in creating the EVMs as explained in Chapter 6.



(a) (b)

Figure 9.2: (a) Unary costs computed from the region descriptors, where the $\mathbf{d}(\cdot,\cdot,\cdot)$ indicates the SIFT descriptor computed on a given patch p, scale σ_p and dominant orientation σ_p. The compared patches are shown as outputs of the operator $\mathbf{d}(\cdot,\cdot,\cdot)$. (b) The proposed pair-wise costs. SIFT descriptors are computed on the affine normalized patches $\overline{q}$ and $\overline{l_q}$ using the dominant gradient orientations θ_p, θ_{l_p} of the regions $\overline{q}$ and $\overline{l_q}$. The compared descriptors $\mathbf{d}(q, \sigma_q, \theta_p)$ and $\mathbf{d}(l_q, \sigma_{l_q}, \theta_{l_p})$ are computed using the texture information of the patches q and l_q, whereas the dominant orientation relative to which the description is performed comes from the neighbouring patches p and l_p. The actual correspondences between the extracted patches are illustrated with different colours, where the same patches with the same colour represent corresponding regions in two different images.

$$V_{pq}(l_p, l_q) = \infty \text{ if } l_p = l_q \neq l_0 \ . \tag{9.4}$$

We further define a *local neighbourhood system* in order to impose flexible local geometric constraints. This neighbourhood system is defined for both the nodes and the labels to impose neighbourhood preservation as the initial geometric constraint (Equation (9.8a)). For regions fulfilling the neighbourhood preservation, the proposed geometric constraint is imposed which measures the consistency of two matches. The local neighbourhood $\mathcal{N}_{\text{MRF}}^{local}(p)$ of a region p is set to be:

$$\mathcal{N}_{\text{MRF}}^{local} = \{q \neq p | \ ||p - q|| < t\} \ , \tag{9.5}$$

where $||p - q||$ is the Euclidean distance between the centres of p and q and t is a threshold value. (We use $t = 10\%$ and $t = 20\%$ of the image size for the node- and label-neighbourhoods respectively to ensure the connectivity of two neighbouring regions after a large viewpoint change).

9.2.3 Pairwise Costs

In order to account for the local geometric relations, we propose a geometric constraint based on the assumption that neighbouring regions move with similar transformations. The idea is as follows: if two neighbouring regions p and q have the corresponding regions l_p and l_q in the second image. Then there exist two affine transformations $\mathbf{A}_p$ and $\mathbf{A}_q$ such that:

$$\begin{aligned} l_p(\mathbf{w}) &= \mathbf{A}_p \cdot p(\mathbf{w}) = s_p \cdot \mathbf{R}_p \cdot \mathbf{M}_p \cdot p(\mathbf{w}) \text{ and} \\ l_q(\mathbf{w}) &= \mathbf{A}_q \cdot q(\mathbf{w}) = s_q \cdot \mathbf{R}_q \cdot \mathbf{M}_q \cdot q(\mathbf{w}) \ , \end{aligned} \tag{9.6}$$

where s_p and s_q are scale factors and $\mathbf{R}_p$ and $\mathbf{R}_q$ are rotation matrices.

Theoretically, for spatially close regions on the same plane it holds $\mathbf{A}_p = \mathbf{A}_q$. However, for neighbouring regions on different planes this assumption is too restrictive and can be relaxed by assuming only $\mathbf{R}_p = \mathbf{R}_q = \mathbf{R}$, where $\mathbf{R}$ is the rotation of the local neighbourhood between two images.

If two neighbouring matches $m_p = (p, l_p)$ and $m_q = (q, l_q)$ are true correspondences, then the SIFT descriptors $\mathbf{d}(\overline{q}, \sigma_{\overline{q}}, \theta_{\overline{q}})$ and $\mathbf{d}(\overline{l_q}, \sigma_{\overline{l_q}}, \theta_{\overline{l_q}})$ computed on the patches $\overline{q}$ and $\overline{l_q}$ using their own characteristic scales $\sigma_{\overline{q}}$, $\sigma_{\overline{l_q}}$ and the dominant gradient orientations $\theta_{\overline{p}}$, $\theta_{\overline{l_p}}$ of the neighbouring regions $\overline{p}$ and $\overline{l_p}$ should be similar. This implies that the two regions within the same local neighbourhood move with similar rotations between two images. The rotation of the neighbourhood can be determined by the relative angle of the dominant gradient orientations $(\theta_{\overline{l_p}} - \theta_{\overline{p}})$ leading to the following rotation matrix:

$$\mathbf{R}(\theta_{\overline{l_p}} - \theta_{\overline{p}}) = \begin{bmatrix} \cos(\theta_{\overline{l_p}} - \theta_{\overline{p}}) & -\sin(\theta_{\overline{l_p}} - \theta_{\overline{p}}) \\ \sin(\theta_{\overline{l_p}} - \theta_{\overline{p}}) & \cos(\theta_{\overline{l_p}} - \theta_{\overline{p}}) \end{bmatrix} \tag{9.7}$$

as illustrated in Figure 9.2(b). (Recall that $\mathbf{M}_p \cdot p = \overline{p}$ and $\mathbf{M}_q \cdot q = \overline{q}$).

This similarity measure indicates the consistency of the matches m_p and m_q, as the rotation estimated from the match m_p is evaluated on the regions of the match m_q. Combining with the neighbourhood preservation, the pairwise costs to evaluate geometric constraints are defined as:

$$\varphi_{pq}(l_p, l_q) = \begin{cases} \infty & \text{if } (l_p \notin \mathcal{N}^{\text{local}}_{\text{MRF}}(l_q)) \quad \text{(9.8a)} \\ \angle\left(\mathbf{d}(\overline{q}, \sigma_{\overline{q}}, \theta_{\overline{p}}), \mathbf{d}(\overline{l_q}, \sigma_{\overline{l_q}}, \theta_{\overline{l_p}})\right) & \text{if } (l_p \in \mathcal{N}^{\text{local}}_{\text{MRF}}(l_q)) \quad \text{(9.8b)} \end{cases}$$

Introducing the geometric constraints, the final pairwise costs are defined as:

$$V_{pq}(l_p, l_q) = \begin{cases} \infty & \text{if } (l_p = l_q \neq l_0) \\ \alpha & \text{if } (l_p = l_0) \wedge (l_q = l_0)) \\ \varphi_{pq}(l_p, l_q) & \text{if } (q \in \mathcal{N}^{\text{local}}_{\text{MRF}}(p)) \\ (p)) & \\ 0 & \text{otherwise} \ , \end{cases} \tag{9.9}$$

where α is the same scale factor used in the unary null-costs in Equation (9.3).

The derived geometric constraint (Equation (9.8b)) is invariant to changes in the scale and relies only on the assumption of similar rotations within a local neighbourhood. This is in contrast to previous methods taking global consistency of the features into account. Such methods would fail in case of global deformation, which is present in our applications. This constraint also allows us to be locally more flexible than recent methods making stronger assumptions, such as invariance of the distance between the neighbouring points or the orientation of line segment connecting their centroids [Zass and Shashua, 2008, Torresani et al., 2008, Leordeanu and Hebert, 2005]. Furthermore, the proposed constraint evaluates local image geometry based on the photometric properties of the patches rather than their spatial locations. This is an important property as it allows for the evaluation of the unary and the pairwise costs in the same space and for their combination in the objective function without using any weighting parameters.

9.2.4 MAP Estimation

After building the MRF model for the deformable wide baseline matching problem, we compute the the maximum a posteriori (MAP) estimate of the optimum labelling l^* using Belief Propagation (BP) [Pearl, 1988].

The non-submodularity of the pairwise costs restricts the choice of the MRF inference algorithms to those without prior constraints on the class of energy functions. Sub-modularity is a property of discrete functions and can be seen as the analogous of convexity of continuous functions [Murota, 2003]. A function φ_{pq} is said to be submodular if it holds:

$$\varphi_{pq}(l_p, l_p) + \varphi_{pq}(l_q, l_q) \leq \varphi_{pq}(l_q, l_p) + \varphi_{pq}(l_p, l_q) \ . \tag{9.10}$$

Due to the uniqueness constraint imposed via the pairwise costs, Equation 9.10 does not hold in our MRF model. Because of its applicability for non-submodular functions, without loss of generality we use the BP algorithm for optimizing the modelled matching problem.

9.3 Evaluation

The performance of the proposed method is evaluated on 4 in-vivo and 4 simulation datasets and compared to 3 matching strategies evaluated in [Mikolajczyk and Schmid, 2005]. The regions are detected and described as explained in Section 9.1. The threshold-based (TB) [Mikolajczyk and Schmid, 2005], nearest-neighbour (NN) [Mikolajczyk and Schmid, 2005] and the nearest neighbour distance ratio matching (NNDR) [Lowe, 2004, Mikolajczyk and Schmid, 2005] matching methods are applied for varying threshold values. The threshold-based (TB) method matches

two regions if the angle between their SIFT descriptors is less than a certain threshold value t:

$$m_{\mathrm{TB}}(p, l_p) = \begin{cases} 1 & \text{if } \angle\left(\mathbf{d}(\overline{p}, \sigma_{\overline{p}}, \theta_{\overline{p}}), \mathbf{d}(\overline{l_p}, \sigma_{\overline{l_p}}, \theta_{\overline{l_p}})\right) < t \\ 0 & \text{otherwise} \quad . \end{cases} \tag{9.11}$$

Note that the TB method can yield multiple matches for the same region.

In the nearest neighbour matching, each region is matched only to its nearest neighbour, if the angle between the two descriptor vectors is smaller then a threshold value t:

$$m_{\mathrm{NN}}(p, l_p) = \begin{cases} 1 & \text{if } (\mathrm{NN}(p) = l_p) \quad \text{and} \quad \angle\left(\mathbf{d}(\overline{p}, \sigma_{\overline{p}}, \theta_{\overline{p}}), \mathbf{d}(\overline{l_p}, \sigma_{\overline{l_p}}, \theta_{\overline{l_p}})\right) < t \\ 0 & \text{otherwise} \quad . \end{cases} \tag{9.12}$$

In the nearest neighbour distance ratio (NNDR) matching proposed by [Lowe, 2004], for each region p in the first image, the first as well as the second nearest neighbours in the second image are found. A match is only accepted if the distance to the second NN is significantly larger than the distance to the first NN:

$$m_{\mathrm{NNDR}}(p, l_p) = \begin{cases} 1 & \text{if } \mathrm{NN}_1(p) = l_p \text{ and } (\mathrm{NN}_2(p)/\mathrm{NN}_1(p)) > t \\ 0 & \text{otherwise} \quad , \end{cases} \tag{9.13}$$

where $\mathrm{NN}_1(p)$ and $\mathrm{NN}_2(p)$ denote the the first and second nearest neighbours of the region p according to the angle between the SIFT descriptors. The ratio $(\mathrm{NN}_2(p)/\mathrm{NN}_1(p))$ can be interpreted as a confidence measure for the established correspondences. Thus, the higher the threshold for this ratio, the more discriminative the matches.

MRF-based method is performed for different values of the factor within the convergence range of the optimization.

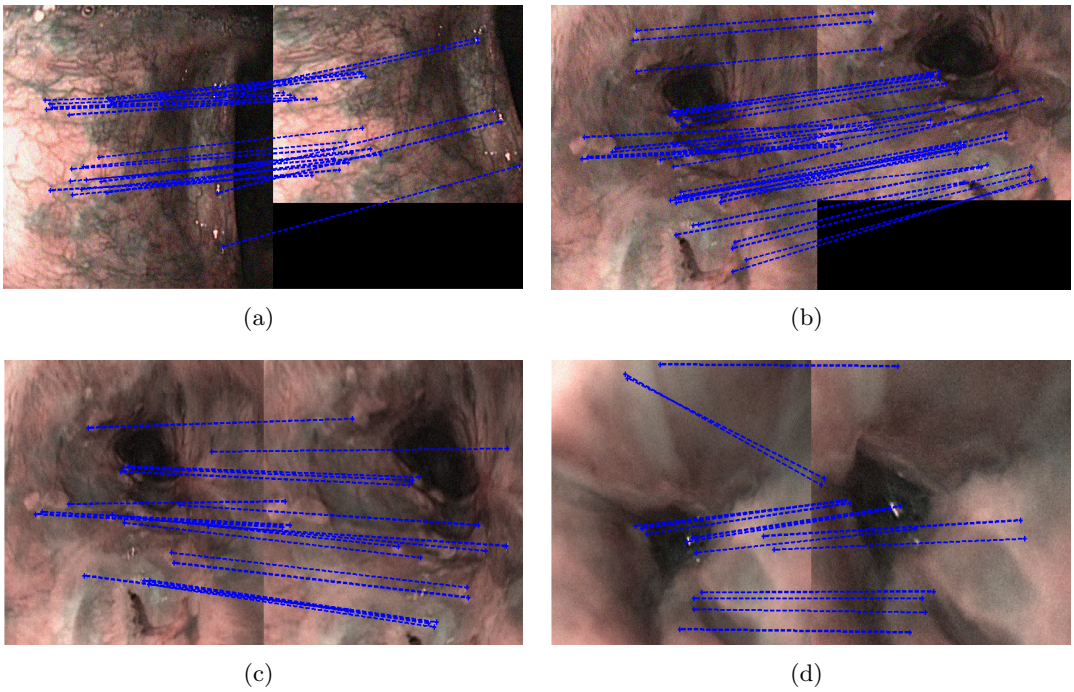
We further compared the hypergraph matching algorithm (HGM) using the proposed affine invariant geometric measure via quadripartite point relations [Zass and Shashua, 2008]. However, these graph matching methods are not adapted to large number of non-matching regions ($43\% - 87\%$ in our datasets). Therefore, the performance of the HGM was not comparable to the other methods and is not illustrated here in detail. For quantitative analysis, we evaluate the F-measure combining the recall[2] and precision[3] values into one quality measure:

$$\mathrm{F}(l^*) = \frac{2 \cdot \mathrm{Precision}(l^*) \cdot \mathrm{Recall}(l^*)}{\mathrm{Precision}(l^*) + \mathrm{Recall}(l^*)} \quad . \tag{9.14}$$

For the best matching results recall=1, precision = 1, and F-measure = 1.

[2]Recall of the matching refers to the ratio of correct matches to the total number of correspondences

[3]Precision of the mathcing denotes the number of true matches with respect to the number of matched regions



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

Figure 9.3: Matching results of the MRF model on simulation datasets with endoscope viewpoint change ((a), (b)) and tissue deformation ((c), (d)).

9.3.1 Simulation Studies

For quantitative evaluation with known ground truth data, we perform two simulation studies. In the first study, we generate images under different viewpoint conditions by transforming 2 in-vivo images from a real patient dataset with known transformations. To this end, we transform two images, one image containing vasculature and one containing anatomical structures, with known homographies. In the second study, we deform 2 in-vivo images, one with anatomical structures and one with largely homogeneous areas and track the detected regions in the deformed frames. The matching results of the MRF model on the simulated datasets are presented in Figure 9.3.

The ground truth information for matching is generated considering the overlap between the detected elliptical regions between within two frames. In both studies, two regions were accepted as a correct match if the distance between the centres of the transformed and detected ellipses was less than 1% of the image size and the overlap was more than 55%. Figures 9.4(a), 9.4(b), 9.4(c) demonstrate that MRF matching results in a better performance than all compared methods for structured scenes. Figure 9.4(d) shows that in the presence of non-distinctive regions, MRF-matching and NNDR (which favours distinctive matches) exhibit a similar performance.

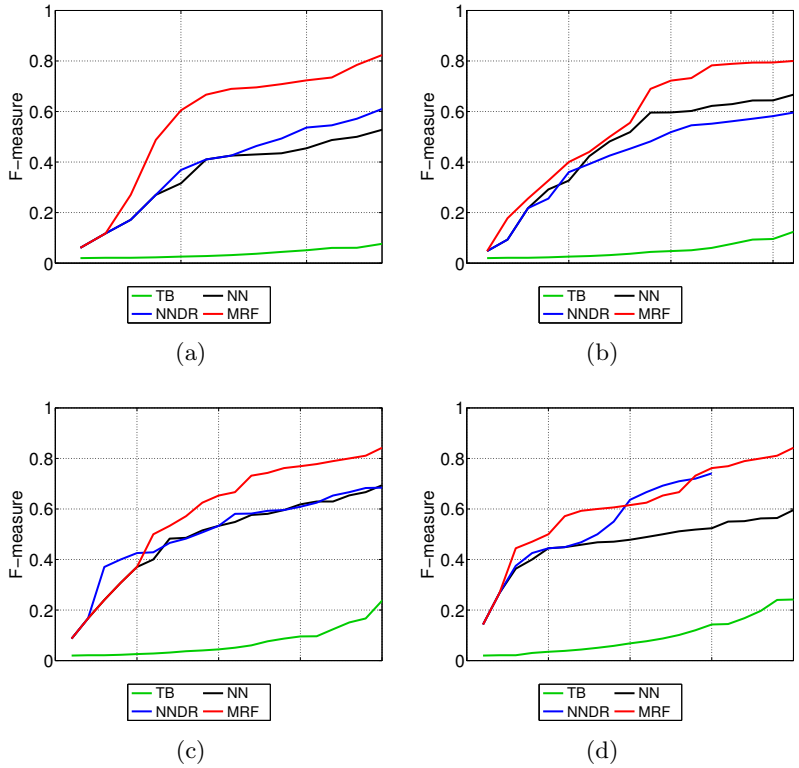


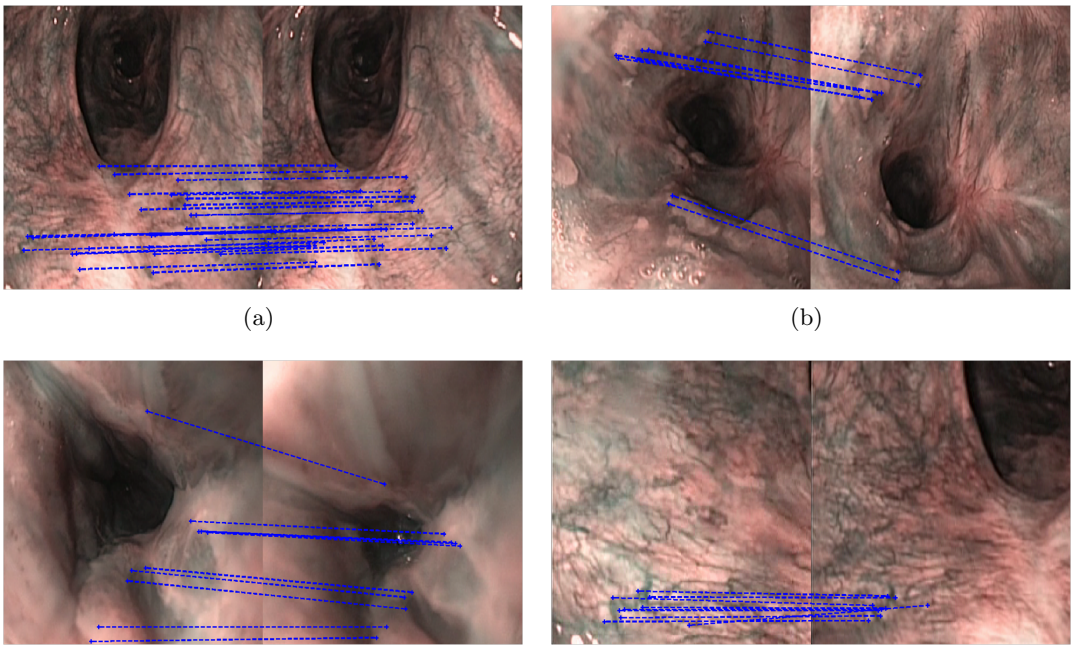
Figure 9.4: (a), (b), (c) and (d) show the F-measure values of each region matching matching algorithm for the 1st, 2nd, 3rd and 4th in-vivo dataset, respectively.

9.3.2 Patient Studies

For the in-vivo studies, we use 4 patient datasets with different viewpoint and photometric conditions. The first 3 datasets contain two distant frames of the same GI procedure from different viewpoints. The frames are chosen to represent encountered endoscopic scenes with different characteristics; i.e. a scene with dense vascular structures (Figure 9.5(a)), a scene with anatomical structures (Figures 9.5(b)) and a scene containing mostly homogeneous areas with large deformation (Figure 9.5(c)). The fourth dataset contains images acquired during two different GI examinations with a time difference of 3 months where the patient underwent chemotherapy (Figure 9.5(a)).

For the in-vivo datasets, the ground truth data is created by manual labelling. Figure 9.6 demonstrates that for all in-vivo cases the proposed MRF model outperforms the compared descriptor matching techniques.

For all datasets (simulation and in-vivo) maximum recall values for the acceptable precision interval (80% – 100% inliers) are summarized in Table 9.1. The matching results for the in-vivo datasets are presented in Figure 9.5.



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

Figure 9.5: Matching results of the MRF model on (a) first, (b) second, (c) third and (d) fourth in-vivo datasets.

Dataset	MRF	NN	NNDR	TB
Viewpoint 1:	**0.75**	0.58	0.53	0.55
Viewpoint 2:	**0.75**	0.66	0.68	0.63
Deformation 1:	**0.76**	0.73	0.62	0.73
Deformation 2:	0.31	0.31	0.31	0.31
In-vivo 1:	**0.96**	0.82	0.75	0.78
In-vivo 2:	**0.69**	0.31	0.31	0.31
In-vivo 3:	**0.71**	0.13	0.06	0.13
In-vivo 4:	**0.83**	0.37	0.21	0.32

Table 9.1: Summary of the maximum recall values for the precision interval [0.8-1.0] (80% – 100% inliers) for the simulation and in-vivo datasets.

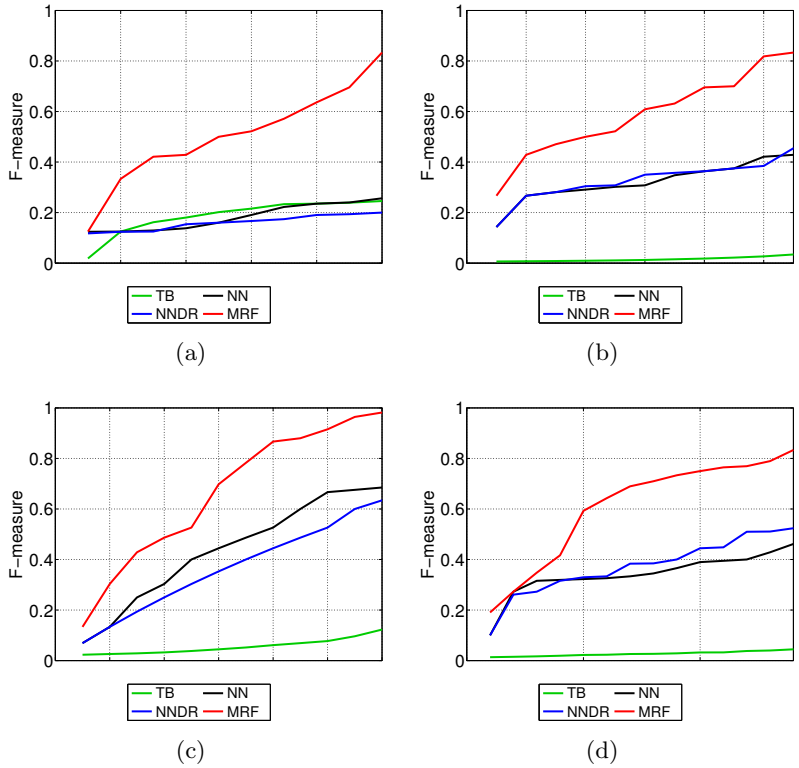


Figure 9.6: F-measure values of all matching algortithms; i.e. threshold-based (TB), nearest-neighbour (NN), nearest-neighbour distance ration (NNDR) and the proposed Markow random field model (MRF) applied to the (a) 1st, (b) 2nd, (c) 3rd and (d) 4th in-vivo dataset.

9.4 Conclusions

In this Chapter, we presented a method for deformable wide baseline matching in order to support the *intra-frame re-targeting* of previous optical biopsy sites. To this end, we proposed an MRF model for matching affine covariant regions. Our model incorporates a novel geometric constraint for dealing with large changes in the observed datasets. Quantitative evaluation presented in Section 9.3 demonstrated the robustness of the proposed model for deformable wide-baseline matching on in-vivo and simulation datasets.

Such intra-frame localisation method provides an alternative for previously discussed point-based localisation methods as presented in [Allain et al., 2009, Allain et al., 2010, Mountney et al., 2009]. In combination with the intra-video localisation method, which is the main focus of this thesis, as presented in Chapters 7 and 8, these intra-frame localisation approaches provide the first attempt towards providing a complete framework for optical biopsy re-targeting in GI endoscopic examinations.

Chapter 10

Conclusions

'If you can't explain it simply, you don't understand it well enough.'
Albert Einstein

In this thesis, we have addressed the task of scene recognition for optical biopsy re-targeting. First, we have analysed the challenges of this re-targeting task, which has recently emerged in the clinical workflow of GI endoscopies due to introduction of the non-invasive optical biopsies. We have identified two separate stages involved in assisting the endoscopic expert in re-targeting the same optical biopsy locations in surveillance endoscopies; i.e. recognition of an optical biopsy scene during the surveillance endoscopy (intra-video localisation) and targeting of the exact biopsy site within the recognized endoscopic frame (intra-frame localisation). In this thesis, we have mainly focused on addressing the intra-video localisation stage but also proposed a complementary intra-frame localisation approach.

In order to address the recognition of optical biopsy scenes in surveillance endoscopies, we have presented a clustering and classification framework. The introduced framework consists of an offline (post-procedural) processing performed on the diagnostic endoscopy and an online (intra-procedural) processing of surveillance endoscopic frames. In the offline processing stage, patient specific endoscopic segments are defined via a two step clustering method. During the online recognition stage, each frame of the surveillance endoscopy is classified as belonging to one of these endoscopic segments. In order to advance our scene recognition to deal with structural changes of the oesophageal tissue, we proposed two run surveillance endoscopies for upper GI examinations. Relying on the proposed workflow for surveillance endoscopies, we extended our individual frame classification by scene correspondences between different examinations.

The framework for optical biopsy re-targeting introduced in this thesis also points to new research directions for future work. The classification schema based on scene

correspondences between two examinations provides an important step towards the daily clinical application of our framework. A further improvement can be achieved by advancing the introduced technique by an automatic cluster matching method. This requires the development of reliable cluster matching techniques in order to establish correspondences between the diagnostic and surveillance videos despite significant structural changes of the oesophageal tissue. Introduction of a reliable cluster matching method into the proposed framework would remove the user-interaction required prior to the second run surveillance endoscopy. Second interesting research direction for future work consists of detecting when the optical probe is introduced into the endoscopic scene during the diagnostic endoscopy. This detection would yield an automatic labelling of optical biopsy scenes of the diagnostic dataset and allow for a seamless integration of the intra-frame localisation into the presented intra-video localisation framework.

The key technical contributions of this thesis are two-fold. Our first contribution lies in the introduction of a new representation for endoscopic video; i.e. the endoscopic video manifolds (EVMs). Thereby, we have pointed out particular mathematical properties of manifold learning framework, which allow for adapting the manifold structure to different applications. As a second contribution, we have introduced two different similarity measures for adapting the manifold stricture to two different clustering tasks. Furthermore, from a theoretical point of view, we have derived new interpretations for the spectral manifold learning methods by drawing on their mathematical equivalence with well-studied physical problems.

Besides the contributions towards addressing the medical task of optical biopsy re-targeting, we have also explored some key properties of human vision as an inspirational source for scene recognition. These explorations resulted in a novel pattern description approach which is presented in the Appendix A of this thesis.

Appendices

Appendix A

A Perceptually Inspired Pattern Description

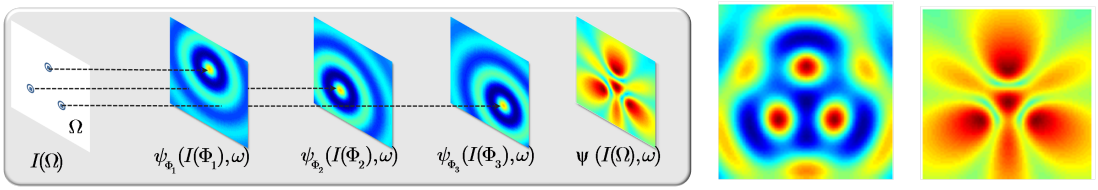
'Science is nothing but perception.'

Plato 428 BC-348 BC

In this appendix, we present a new pattern description method inspired by the study of human perception. In Chapter 2, we have discussed our initial study on expert's perception of endoscopic images. Following this research direction further, we perform a more detailed study on some key properties of human perception. Inspired by three important properties of perception as pointed out by the Gestalt theory [Koffka, 1999], we present a new pattern description method based on wave interference. In the rest of this appendix, we refer to the presented pattern description approach as *"interference description"* (ID). Furthermore, we demonstrate that the interference phenomenon intrinsically exhibits some of the key Gestalt properties seen in human perception.

A.1 Properties of Human Perception

For many years, human vision has been an inspiring model for solving many different computer vision tasks such as feature detection, object recognition, image representation, and learning. Several principles of human perception, such as invariance, emergence and reification (dynamic grouping) have been identified and studied extensively in Gestalt psychology [Koffka, 1999]. Substantial work has been carried out in applying the two properties, invariance and dynamic grouping, for several computer vision tasks [Lowe, 1999, Serre et al., 2007, Dollar et al., 2006, Dubuc and Zucker, 2001a, Dubuc and Zucker, 2001b, Geisler et al., 2001, Williams and Jacobs, 1997,



(a) (b) (c)

Figure A.1: (a) Demonstration of the emergent nature of interference. Each source in the stimulus field creates a (attenuating) circular wave on the medium resulting in the final interference pattern. (b) Sum of the amplitudes of individual wave patterns. (c) Interference of individual wave patterns. The interference phenomena intrinsically exhibits the emergence property, as the whole (the interference pattern) shown in (c) is different than the sum of its constituent wave profiles shown in (b).

Guo et al., 2003, Ren et al., 2005, Desolneux et al., 2007]. However, there exists another property, called *emergence*, which implies that an object is not recognized by first perceiving its parts and combining them to a meaningful whole but rather as an emergent pattern; i.e. *"The whole is different than the sum of the parts"* [Koffka, 1999].

In fact, in nature there exists a physical phenomenon, the interference of waves, that intrinsically exhibits the emergence property. The interference pattern is produced by the superposition of two or more waves. This results in occurrence of new structures that are not present in the constituent waves but are determined by their interrelations as shown in Figure A.1(a). Indeed, a property is called *emergent* if it is not present in any of the constituent parts but only in their combination. For instance, the interference pattern created by three spherical waves (Figure A.1(c)) contains structures, which are not present in any of the constituent waves (Figure A.1(a)) nor in their simple sum (Figure A.1(b)).

As next we discuss three of the key Gestalt properties involved in our pattern description.

A.1.1 Invariance

Invariance is probably the most relevant argument of Gestalt theory and concerns the ability of human observers to recognize patterns and objects independent of their location, scale rotation and even deformation as illustrated in Figure A.2(a). In computer vision literature, considerable effort has been directed toward achieving invariance to rotation, change in scale and illumination for object matching and recognition. Several methods address this problem by relying on the detection and the description of distinctive local image features without accounting for the geometry of the extracted features (Examples are [Lowe, 1999, Lowe, 2004, Sivic and Zisserman, 2009]). An interesting approach inspired by the biology of the visual cortex is proposed in [Serre et al., 2007]. In this work, the authors present a hierarchical framework for object recognition and

categorization which is based on the organization of visual cortex. The proposed method alternates between convolution and maximum operators in order to create increasingly invariant representations of the input image at each stage of the hierarchical system.

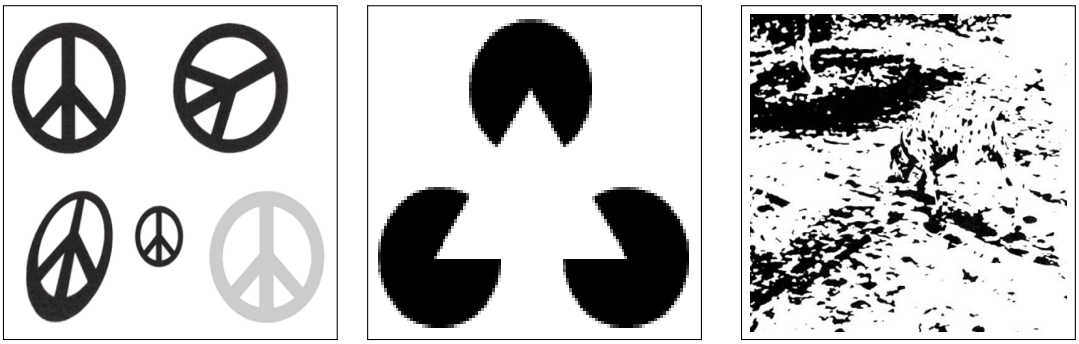
A.1.2 Reification

Reification is the constructive property seen in human perception. In Figure A.11(a), a triangle is perceived although it is not explicitly delineated. This is due to the spatial configuration of the three packman shapes. The reification argument is also related to the dynamic grouping property of Gestalt theory, which states that the information from two or more sources in stimuli (image) is combined dynamically to create a joint description [Watt and Phillips, 2000]. Studies have proposed mechanisms to account for this property [Williams and Jacobs, 1997, Elder and Goldberg, 2002, Dubuc and Zucker, 2001a, Dubuc and Zucker, 2001b]. Elder and Goldberg construct a Bayesian model for contour grouping [Elder and Goldberg, 2002], Williams and Jacobs present a contour completion approach using random walks [Williams and Jacobs, 1997]. Dubuc and Zucker propose a contour grouping and segmentation approach based on measuring representational complexity [Dubuc and Zucker, 2001a, Dubuc and Zucker, 2001b]. Dollar *et al.* present a learning based algorithm for object boundary grouping which also satisfies the Gestalt grouping argument [Dollar et al., 2006]. Finally, Bileschi and Wolf create intermediate image features based on the dynamic grouping property of the Gestalt theory [Bileschi and Wolf, 2007].

A.1.3 Emergence

Besides these two properties, which have attracted attention in the computer vision community, there exists another key argument of Gestalt theory, namely the emergence. *Emergence* property states that patterns are recognized in an emergent manner, where the information in the complete pattern is different than the sum of the constituent parts. A property is called *emergent* if it is not present in any of the constituent parts but only in their combination. An example can be seen in Figure A.11(b). The Dalmatian in the well known photograph by R. C. James[1] is not perceived by first detecting the parts but rather emerges as a whole pattern due to the presence and interrelations of the parts. In an emergent mechanism there exists no rule describing the global property but it is formed due to the interactions of the parts. Although there exit several description methods considering the contextual relations, such as [Belongie et al., 2002], this is usually done by a set of defined rules. We model the emergence by considering that it is inherent to wave interference.

[1]The famous photograph by R. C. James captures a Dalmatian dog on a cluttered background. A manual segmentation of the Dalmatian for better illustration is presented in Figure A.11.



(a) (b) (c)

Figure A.2: (a) Invariance b) reification and (c) emergence properties seen in human perception as pointed out by the Gestalt theory. (a) Human observers are able to recognize objects and patterns independent of their location, scale and rotation. (b) Reification refers to the constructive property seen in human perception. Observers perceive a triangle in this image, although it is not explicitly delineated. (c) Emergence property states that objects are not recognized by first detecting individual parts and then combining them to a meaningful whole but rather in emergent manner. The Dalmatian dog, photographed by R. C. James, in this image is not recognized by first detecting its parts, as many of the parts are not even visible in the picture but rather emerges as a whole.

A.2 Properties of Wave Interference

In its most basic form, a wave can be defined as a pattern (or distribution) of disturbance that propagates through a medium as time evolves. Harmonic waves are waves with the simplest wave profile, namely sine or cosine functions. Due to their special waveform they create wave patterns which are not only periodic in time but also in space. Thus, on a 2D medium, harmonic waves give rise to *circular wave patterns* such as the ones observed on a liquid surface when a particle is dropped into the liquid.

Using complex representation, a circular wave is described by the following wave function of any point $\mathbf{x}$ on the medium and a time instance t:

$$\psi_{\boldsymbol{\phi}}(A_0, \omega; \mathbf{x}, t) = A_0 \cdot e^{i(\omega||\mathbf{x}-\boldsymbol{\phi}||)} \cdot e^{i\omega t} \quad , \tag{A.1}$$

where i is the imaginary unit, ω is the frequency of the wave and $\boldsymbol{\phi}$ is the location of the source (stimulus). A_0 denotes the intensity of the source and $||\mathbf{x} - \boldsymbol{\phi}||$ gives the distance of each point $\mathbf{x}$ on the medium to the location of the source of disturbance.

As the time dependent term $e^{i\omega t}$ is not a function of space ($\mathbf{x}$), it does not affect the interference of the waves. Therefore, in the rest of this appendix it will be taken as $t = 0$. So, the instantaneous amplitude of the disturbance at each point on the medium is given

by the real part of the wave function Real $[\psi_{\boldsymbol{\phi}}(A_0, \omega; \mathbf{x})]$. For simplicity, we will denote waves with a fixed frequency ω and a fixed amplitude A_0 as $\psi_\phi(\mathbf{x}) := \psi_\phi(A_0, \omega; \mathbf{x})$.

As waves propagate outwards from the source, the amplitude of the wave will gradually decrease with the increased distance from the source due to friction. This effect, called *attenuation*, can be described by multiplying the amplitude of the wave with the attenuation profile σ.

When two or more waves ψ_{ϕ_i} are created on the same medium at the same time from several source locations ϕ_i, the resultant disturbance Ψ at any point $\mathbf{x}$ on the medium is the algebraic sum of the amplitudes of all constituents. Using complex representation, the interference pattern $\Psi(\mathbf{x})$ is given by the amplitude of the sum of individual complex wave functions $\psi_{\phi_i}(\mathbf{x})$:

$$\Psi(\mathbf{x}) = \left| \sum_{i=1}^{n} \psi_{\phi_i}(\mathbf{x}) \right| \quad , \tag{A.2}$$

where $|\cdot|$ denotes the absolute value of the complex number. Note that the absolute value of the sum of a set of complex numbers is different than the sum of their real parts (See Figure A.1). This is known as the superposition principle of the waves. The superposition of waves with fixed frequencies yields to a special phenomenon called *interference*. In this special case, if the waves reaching a point on the medium are in-phase (aligned in their ascent and descents), they will amplify each other's amplitudes (constructive interference); conversely if they are out-of-phase, they will diminish each other at that point (destructive interference). This results in a new wave profile called interference pattern.

Figure A.1(a) illustrates the interference pattern created by the superposition of 3 waves with the same frequency. The interference pattern is computed by the absolute value of the sum of complex wave functions. The complex addition results in a waveform which is not present in any of the constituting waves nor in the simple sum of the constituent wave amplitudes (Figure A.1(b)) but occurs due to their interference (Figure A.1(c)). Therefore, interference property of waves provides an emergent mechanism for describing the relations between the sources and for propagating local information.

A.3 Interference Description

Using the introduced definitions in Section A.2, the presented pattern description method formulates the contextual relations of the parts composing an image in an emergent manner.

Let $\Omega \subset \mathbb{R}^2$ be the medium on which the waves are created, where $\mathbf{x} \in \Omega$ denotes a point on Ω. In practice the medium is described with a grid discretising the space. The function $I : \Omega \to \mathbb{R}$, called stimulus field, assigns a source intensity $A_0 = I(\phi)$ to the positions ϕ on the medium Ω. This function contains the local information of the

pattern which is propagated over the whole medium. Depending on the application, it can be chosen to be a particular cue defined on the image domain such as gradient magnitude, extracted edges or features.

Each value $I(\phi)$ of the stimulus field I induces a circular wave $\psi_{\boldsymbol{\phi}}(I(\boldsymbol{\phi}), \omega; \mathbf{x})$ on the medium. We use the frequency of the wave to describe the relation between the distribution of sources and the scale at which one wants to describe the pattern. This is discussed in detail in Section A.3.1. The profile of the created circular wave is described as:

$$\psi_{\boldsymbol{\phi}}(I(\boldsymbol{\phi}), \omega; \mathbf{x}) = I(\boldsymbol{\phi}) \cdot e^{i(\omega(||\mathbf{x}-\boldsymbol{\phi}||)} \cdot \sigma \ , \tag{A.3}$$

where σ is the attenuation profile. We define the attenuation profile of the wave induced by the source ϕ with frequency ω as:

$$\sigma(\boldsymbol{\phi}, \omega; \mathbf{x}) = \omega \cdot e^{-\omega \cdot ||\mathbf{x}-\boldsymbol{\phi}||} \ , \tag{A.4}$$

where $||\mathbf{x}-\boldsymbol{\phi}||$ denotes the Euclidean distance between a point $\mathbf{x}$ and the source location $\boldsymbol{\phi}$. This attenuation profile allows us to conserve the total energy of the waves while changing their frequency. Note that the attenuation profile is a function of the distance to the source ($||\mathbf{x} - \boldsymbol{\phi}||$) as well as the frequency ω. The higher the frequency, the sharper is the decay of the attenuation profile. This yields to a localized effect for high frequencies and a more spread effect for low frequencies.

Once the location and intensities of the sources on the stimulus field are determined, the interference pattern $\Psi(I, \omega; \mathbf{x})$ on the medium for a fixed frequency ω is computed simply by the superposition of all wave patterns:

$$\Psi(I, \omega; \mathbf{x}) = \left| \sum_{\phi_i \in \Omega} \psi_{\phi_i}(I(\phi_i), \omega; \mathbf{x}) \right| \ . \tag{A.5}$$

A.3.1 Multi-Frequency Analysis

Computing the interference patterns for several frequencies allows for the analysis of the stimulus field at different scales. As discussed in Section A.3, high frequency waves are more localized and propagate the content in a smaller neighbourhood, whereas low frequency waves are more spread over a larger area. Figure A.3 illustrates the circular wave patterns for several frequencies and source locations.

Formally, we define the interference description (ID) $\Theta(I)$ of an input I as the set of interference patterns $\Psi(I, \omega_i)$ for a range of frequencies $\{\omega_1, \omega_2, ..., \omega_n\}$:

$$\Theta(I) = \{\Psi(I, \omega_1), \Psi(I, \omega_2), \cdots, \Psi(I, \omega_n)\} \ , \tag{A.6}$$

with $\omega_i = \frac{2\pi \cdot k_i}{P}$, where $k_i \in \mathbb{N}$ is the wave-number, i.e. the number of periods the wave has on the medium and P is the length of the grid.

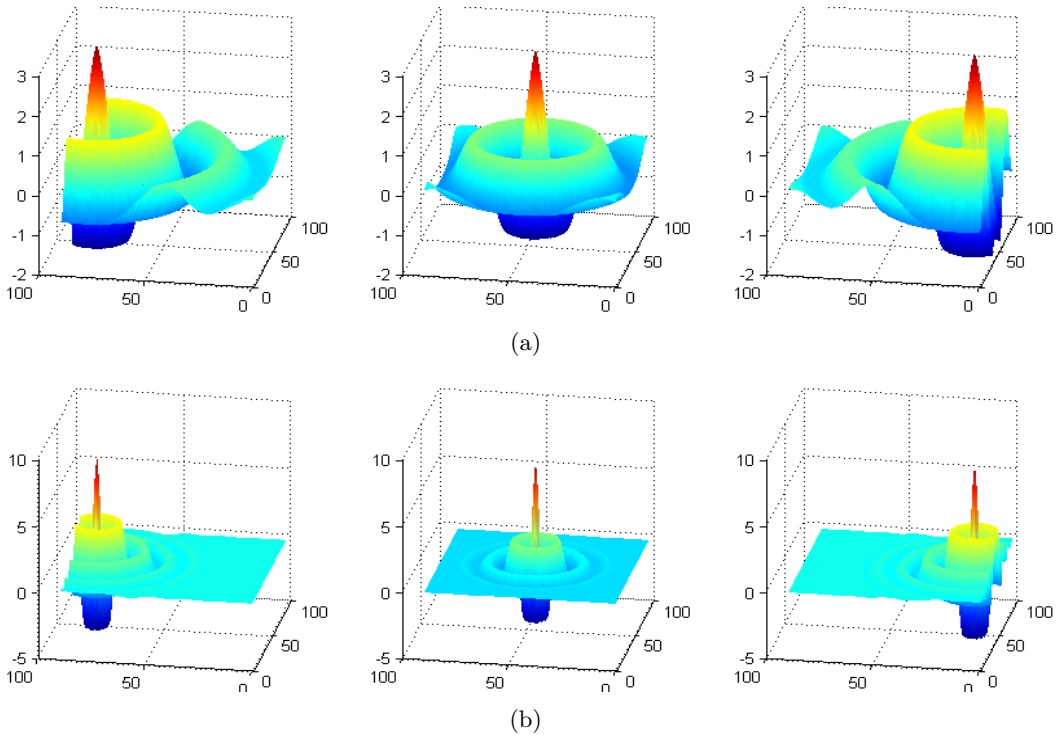


Figure A.3: Wave patterns for different source locations with the 3rd frequency ($k = 3$) shown in a) and 8th frequency ($k = 8$) shown in b).

A.3.2 Descriptor Comparison

In this section, we present a method to compare two IDs, $\Theta(I_1)$ and $\Theta(I_2)$ in a rotation invariant manner. To this end, we first define a *coherency measure* c of an interference pattern $\Psi(\mathbf{x})$:

$$c(\Psi(\mathbf{x})) = \frac{|\Psi(\mathbf{x})|}{\sum_{\phi_i \in \Omega} |\psi_{\boldsymbol{\phi}_\mathbf{i}}(\mathbf{x})|} \ . \tag{A.7}$$

The coherency measures the power of the interference pattern compared to the sum of the powers of the constituent waves. As $\sum_{\phi_i \in \Omega} |\psi_{\boldsymbol{\phi}_\mathbf{i}}(\mathbf{x})|$ provides the upper limit to the power of the interference amplitude $|\Psi(\mathbf{x})|$ for each location $\mathbf{x}$ on the medium, the coherency measure c normalizes the values of the interference profile to the interval $[0, 1]$ (Figure A.4(c), A.4(d)). This enables the comparison of different IDs independent of their absolute values and therefore makes the comparison *independent of the number of sources*.

We subdivide this coherency image $c(\Psi(\mathbf{x}))$ into $m + 1$ levelsets

$$\{\gamma_0, \gamma_1, \cdots, \gamma_j, \cdots, \gamma_m\} \ , \quad \gamma_j = \{\mathbf{x} | c(\Psi(\mathbf{x})) = j/m\} \ , \tag{A.8}$$

and take the sum of the coherency values for each interval between the two consecutive

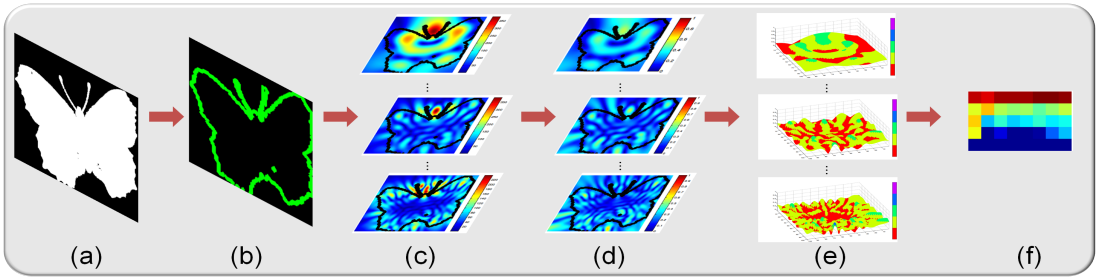


Figure A.4: Rotation invariant multi-frequency descriptor. a) Input image, b) sources located on the medium Ω as marked in green, c) interference patterns computed for a range of frequencies $\Theta(I)$, d) coherence measure computed on the interference patterns $c(\Theta(I))$, e) Levelsets $\{\gamma_1, \cdots, \gamma_m\}$ and f) the levelset histogram, where the columns correspond to descriptors at different frequencies.

levelsets. This yields an m-dimensional measure

$$\mathbf{h} = [h1, \cdots, h_m]^\mathsf{T} \ , \quad h_j = \sum_{\gamma_{j-1} \leq x < \gamma_j} c(\Psi(\mathbf{x})) \tag{A.9}$$

computed from an interference pattern $\Psi(\mathbf{x})$ (Figure A.4(e)). Performing this levelset histogram calculation for the whole range of frequencies n of an ID ($\Theta(I)$) results in a $m \times n$ dimensional measurement:

$$\mathbf{h}(\Theta(I)) = [\mathbf{h}(\Psi(I, \omega_1)), \cdots, \mathbf{h}(\Psi(I, \omega_n))] \tag{A.10}$$

as illustrated in Figure A.4(f). Quantifying the amount of coherency in each levelset enables a rotation-invariant description of the created interference patterns. The distance between the two IDs $\Theta(I_1)$ and $\Theta(I_2)$ is then simply computed as the angle between the levelset histograms:

$$dist(I_1, I_2) = \arccos\left(\frac{\langle \mathbf{h}(\Theta(I_1))^\mathsf{T}, \mathbf{h}(\Theta(I_2))\rangle}{||\mathbf{h}(\Theta(I_1))|| \cdot ||\mathbf{h}(\Theta(I_2))||}\right) \ . \tag{A.11}$$

This dissimilarity measures the distance between the two patterns in a rotation invariant manner (due to the levelset histograms) while taking into account the contextual relations between the constituent parts (due to the interference pattern) for each pattern.

A.3.3 Applications and Evaluation

Firstly, we demonstrate that the ID exhibits the Gestalt properties discussed in Section A.1. Then, we show its application for shape matching and retrieval and pattern recognition. For our experiments, we first create a grid (medium $\Omega : [-\pi \times \pi] \times [-\pi \times \pi]$) and map the locations of the stimulus field to the sources on the medium. Each source creates a circular wave pattern propagating outwards from this source location. Also, we define the stimulus field to be the gradient magnitude of the input image.

A.3.3.1 Verification of Gestalt Properties

Firstly, we demonstrate how the proposed description exhibits the Gestalt properties emergence, reification and invariance.

Emergence:
The emergence property states that the whole (structure) is different than the sum of its parts. This is due to the interrelations of the constituents. In ID, the whole (*i.e.* interference pattern as shown in Figure A.1(c)) is given by the amplitude of the complex interference field (sum of the complex wave functions). This interference pattern is different from the sum of the amplitudes of the constituent waves (Figure A.1(b)). This is due to the constructive and destructive interference effect (as explained in Section A.2) and can be seen mathematically as:

$$\Psi(\mathbf{x}) = \left| \sum_{i=1}^{n} \psi_{\phi_i}(\mathbf{x}) \right| \neq \sum_{i=1}^{n} \text{Real}[\psi_{\phi_i}(\mathbf{x})] = \sum_{i=1}^{n} \psi_{\phi_i}(\mathbf{x}) \ , \tag{A.12}$$

where the left side of the inequality defines the interference pattern and the right side gives the sum of the amplitudes of its constituent waves. The difference between the whole and the sum of its parts leads to the emergence of a new profile which is not present in any of the constituent circular waves nor in their simple sum but occurs due to their interrelations.

Reification:
Reification is the constructive property seen in human perception [Koffka, 1999, Carman and Welch, 1992]. In Figures A.5(b) and A.5(f), a triangle and a square are perceived, although they are not explicitly delineated. This is due to the particular spatial arrangement of the packman shapes. ID results in emergence of very similar interference patterns in the location where a triangle is perceived in Figures A.5(a) and A.5(b) and where a square is perceived in Figures A.5(e) and A.5(f). Therefore, ID also allows to recognize patterns with missing contours such as the triangle image in Figure A.10(d).

Invariance:
The invariance of ID to scale change is achieved due to the mapping of the stimulus field onto the same medium (discretized grid of fixed size). This creates sources along contours of the same size, however, the image with a smaller scale results in less number of sources sampled along the contour. The structure of the created interference patterns are the same as shown in Figures A.6(h), A.6(i). As the proposed comparison method in Section A.3.2 is independent of the number of sources, the two IDs result in very similar levelset histograms as shown in Figure A.6(o), A.6(p).

We further demonstrate the invariance of the ID for rotation, affine and perspective transformations, as well as deformation. To this end, we simulate a binary triangle image and apply the corresponding transformations. Figures A.6(c)-A.6(g) display for each transformation, the input images (Figure A.6(c)-A.6(g)), interference patterns for the same example frequency (Figure A.6(j)-A.6(n)) and the corresponding levelset

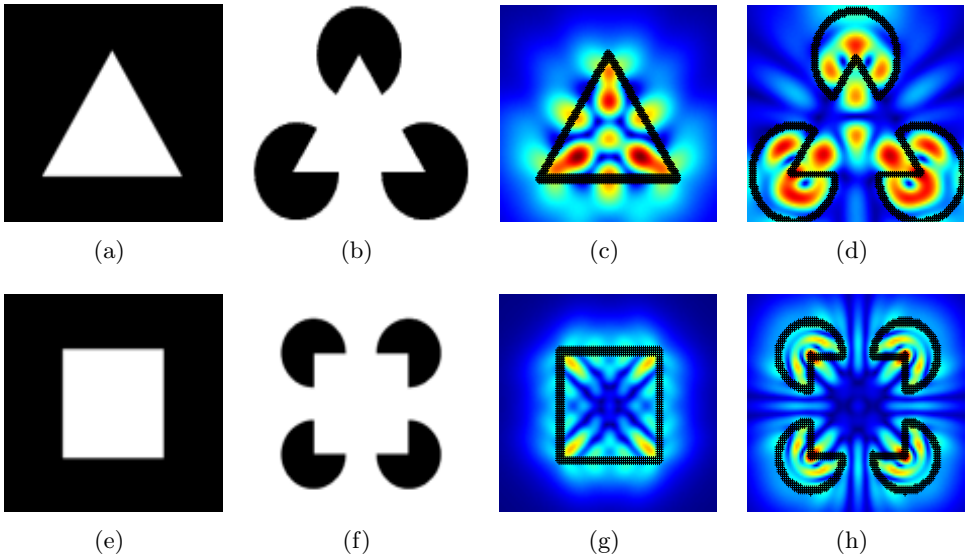


Figure A.5: Reification property seen in wave interference. Similar interference patterns emerge for a complete triangle a) and a Kanizsa triangle image b) or a complete square e) and a Kanizsa square image f). c) and d) demonstrate the interference patterns for the 4th frequency ($k = 4$) for the triangle and Kanizsa triangle, respectively. The source locations are illustrated in black crosses. g) and h) show the interference patterns of images in e) and f) for the 8th frequency ($k = 8$).

histograms (Figure A.6(q)-A.6(u)), computed as described in Section A.3.2. Note the similarity of the IDs and the levelset histograms despite the large geometric transformations between the images.

A.3.3.2 Application to Shape Matching and Retrieval

In this experiment, we demonstrate the use of IDs for shape matching. We apply the presented description approach to match objects from the contour images of the MPEG7 CE shape database containing 1400 images with 70 different categories[2]. For each image, we first compute the interference patterns for 10 frequencies. Then we compute the similarity of each shape image against all other images in the database by comparing the IDs as described in Section A.3.2. Figure A.7 demonstrates the performance of ID matching as a confusion matrix computed from the complete MPEG7 database.

After matching the shape images, we assign the best matched category to each image for retrieval considering the 20 minimum distances. Figure A.8 displays the recall and precision over the 70 categories.

[2]The images can be downloaded from `http://www.imageprocessingplace.com/downloads_V3/root_downloads/image_databases/MPEG7_CE-Shape-1_Part_B.zip`.

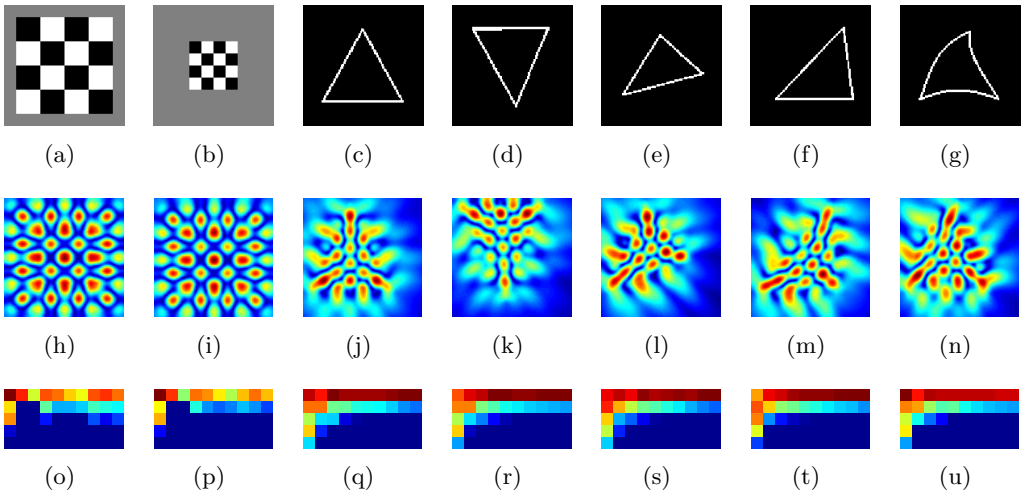


Figure A.6: Invariance of ID for scale change (a,b), rotation (c,d), affine transformation (c,e), perspective transformation (c,f) and deformation (c,g). First row shows the input patterns, second row illustrates interference pattern for the 6th frequency $k = 6$ and last row demonstrates the levelset histograms computed from the IDs $h(\Theta(I))$.

A.3.4 Pattern Recognition

In this final experiment we demonstrate that ID can be used to recognize patterns created by different texture elements. Therefore, we use images with large inter-class variability in the representation where in each image the pattern is generated with different texture elements. Although the details of the patterns are different, they all share the same global layout. This global structure is enhanced in the low frequency interference patterns. Therefore we create IDs $\Theta(I)$ for $k = \{1, \cdots, 4\}$.

Figures A.9 and A.10 illustrate IDs created from the input images with similar layout but different texture elements. IDs lead to similar but distinctive patterns despite the differences in the texture elements (parts) creating the global pattern (the whole). Note that Figures A.10(a)-A.10(e) yield to more similar interference patterns compared to Figures A.10(k)-A.10(o), although the global layout is shared within both classes (Figures A.10(a)-A.10(e) and Figures A.10(k)-A.10(o)). This is also in agreement with the perception of human observers for the similarity of these patterns.

A.4 Conclusions

In this appendix, we presented a method for pattern description that is based on wave interference. Due to the characteristics of interference phenomenon, our method intrinsically accounts for the contextual relations between the parts of a pattern. This eliminates the need for defining a set of rules or for any prior learning. Furthermore, the

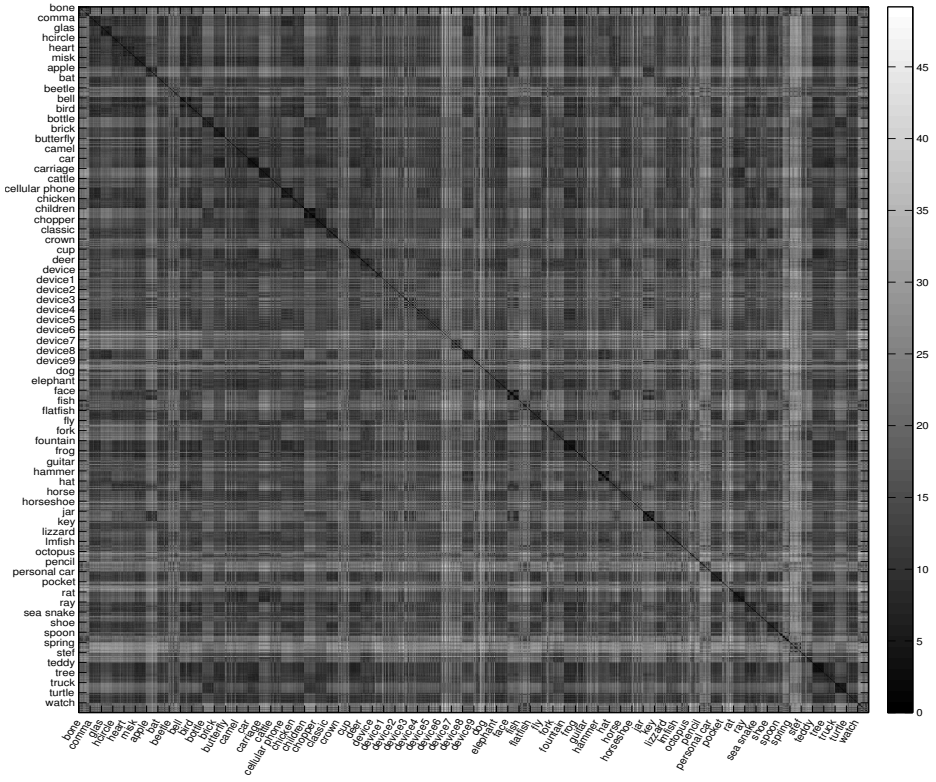


Figure A.7: The confusion matrix of ID matching as applied on the MPEG7 database.

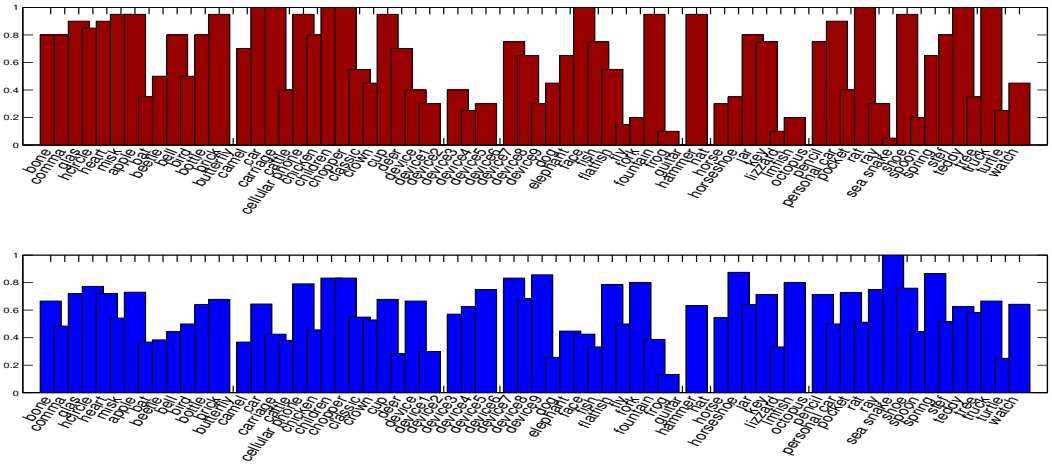


Figure A.8: a) Recall and b) precision values for shape category retrieval of IDs when applied to all categories in the MPEG7 shape database.

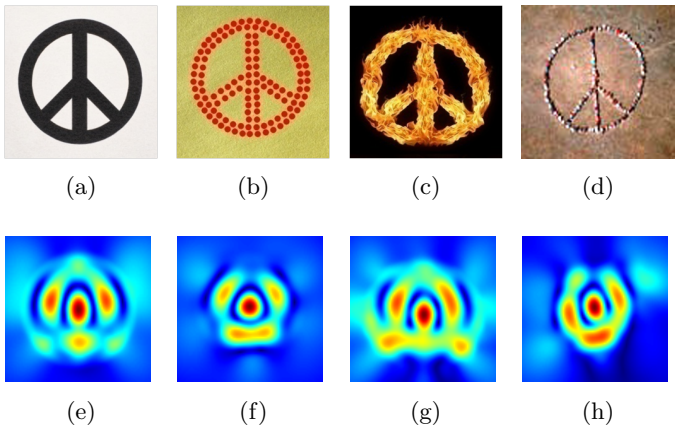


Figure A.9: Application of ID for pattern recognition. (a), (b), (c) and (d) show the input images containing the same global layout created with different texture elements. (e)-(h) IDs at the frequency ($k = 3$) of the images (a-(d).

particular mathematical formulations of the presented method allow for the computation of higher order relations in an efficient manner, i.e. a simple complex addition.

The proposed mathematical model also allows for further improvement via two simple modifications. Firstly, the interference of only selected features (instead of the whole stimuli) can be considered at different frequencies; i.e. more global features can be defined to interfere at low frequencies, whereas more specific information is propagated using the high frequencies. Secondly, in this work, all source points (stimuli) create waves with the same phase; i.e. all circular waves are created at the same time. However, the particular mathematical model allows for including further information, if desired, simply by changing the initial phase of the features (sources).

In a series of experiments, we demonstrated that the proposed description is in agreement with three key Gestalt properties relevant for pattern description and pointed out the analogies between these properties and the wave interference phenomenon.

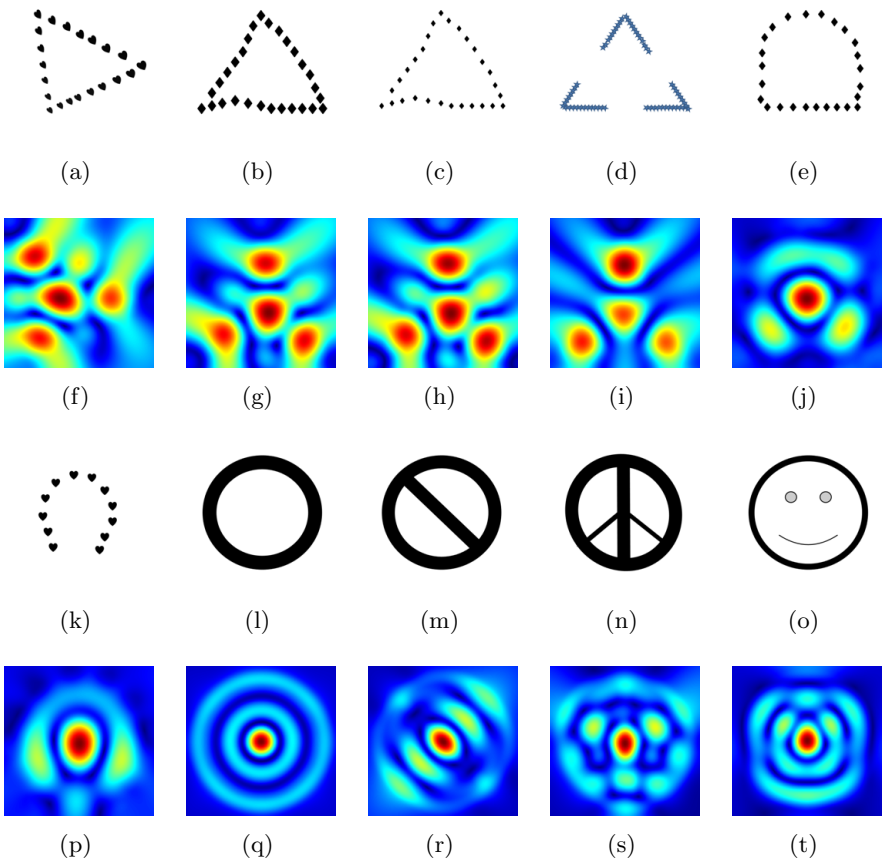


Figure A.10: Interference patterns created from input images (shown in (f)-(i)) with several geometric transformations ((a)-(c)) and missing contours (d) leads to the emergence of similar patterns (shown in (a)-(d)), whereas shapes with only very small variations (shown in (k)-(o)) yield very distinct interference patterns (shown in (p)-(t)).

(a)

(b)

Figure A.11: (a) Well known picture of a dalmatian photographed by R. C. James. b) The same photograph in (a) where the dalmatian is manually marked with ref for illustrated. Note the absence of several parts of the dalmatian in the original image. This leads to the interpretation that in human vision the objects are not recognised by first extracting the parts and then combining them into a meaningful whole but the object is perceived as a whole in rather emergent manner; i.e. the whole is different than the sum of its parts.

Appendix B

List of Abbrevations

B.1 Medical Terms

BO	Barrett's Oesophagus
CT	Computed Tomography
$\mathcal{D}$	Diagnostic (Endoscopy)
GI	Gastro-intestinal
MRI	Magnetic Resonance Imaging
NBI	Narrow-Band Endoscopic Imaging
OAC	Oesophageal Adenocarcinoma
pCLE	Probe-based Confocal Endomicroscopy
PSES	Patient Specific Endoscopic Segments
$\mathcal{S}1$	First Run Surveillance (Endoscopy)
$\mathcal{S}2$	Second Run Surveillance (Endoscopy)
WLE	White Light Endoscopy

B.2 Technical Terms

a.e.	almost everywhere
Bhat.	Bhattacharyya
BCD	Between Cluster Distances
BP	Belief Propagation
CCD	Charge-Coupled Device
Cos.	Cosine
CD-measure	Compactness and Separability measure
D	Dimensional
BD-index	Davis-Bouldin index
DoF	Degrees of Freedom
EH	Energy Histogram
ED	Euclidean Distance
EDst	Energy Distribution
EVM	Endoscopic Video Manifold
FMM	Finite Mixture Models
fp	false positive
HGM	Hyper-Graph Matching
ID	Interference Description
LE	Laplacian Eigenmaps
LLE	Locally Linear Embedding
LOPO	Leave-One-Part-Out
LPP	Locality Preserving Projections
NCC	Normalized Cross Correlation
NN	Nearest Neighbour
NNDR	Nearest Neighbour Distance Ratio
MAP	Maximum a posteriori
MRF	Markov Random Field
PCA	Principal Components Analysis
PDE	Partial Differential Equation
RP	Reproducing Kernel
RPHS	Reproducing Kernel Hilbert Space
SIFT	Scale Invariant Feature Transform
SSD	Sum of Squared Distances
TB	Threshold Based
tp	true positive
vtEVM	visual temporal Endoscopic Video Manifold
WCD	Within Cluster Distances

Appendix C

List of Publications

Journals Publications

- Atasoy, S., Mateus, D., Meining, A., Yang, G.-Z., and Navab, N. (2012). Endoscopic video manifolds for targeted optical biopsy. *IEEE Transactions on Medical Imaging*, 31:637–653.

- Meining, A., Atasoy, S., Chung, A., Navab, N., and Yang, G. (2010). "eye-tracking" for assessment of image perception in gastrointestinal endoscopy with narrow-band imaging compared with white-light endoscopy. *Endoscopy*, 42(8):652–655.

Peer-Reviewed Conference Publications

- Atasoy, S., Mateus, D., Meining, A., Yang, G.-Z., and Navab, N. (2011). Targeted optical biopsies for surveillance endoscopies. In Fichtinger, G., Martel, A., and Peters, T., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 6893 of *Lecture Notes in Computer Science*, pages 83–90. Springer Berlin / Heidelberg.

- Atasoy, S., Mateus, D., Georgiou, A., Navab, N., and Yang, G.-Z. (2010a). Wave interference for pattern description. In Kimmel, R., Klette, R., and Sugimoto, A., editors, *Proceedings of Asian Conference on Computer Vision (ACCV)*, volume 6493 of *Lecture Notes in Computer Science*, pages 41–54. Springer Berlin / Heidelberg.

- Atasoy, S., Mateus, D., Lallemand, J., Meining, A., Yang, G.-Z., and Navab, N. (2010b). Endoscopic video manifolds. In Jiang, T., Navab, N., Pluim, J., and Viergever, M., editors, *Proceedings of International Conference on Medical*

Image Computing and Computer Assisted Intervention (MICCAI), volume 6362 of *Lecture Notes in Computer Science*, pages 437–445. Springer Berlin / Heidelberg.

- Meining, A., Atasoy, S., Navab, N., Chung, A. J., and Yang, G.-Z. (2009). Targeted optical biopsies for surveillance endoscopies. In *GASTRO*, London, UK.

- Atasoy, S., Glocker, B., Giannarou, S., Mateus, D., Meining, A., Yang, G.-Z., and Navab, N. (2009). Probabilistic region matching in narrow-band endoscopy for targeted optical biopsy. In Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., and Taylor, C., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 5761 of *Lecture Notes in Computer Science*, pages 499–506. Springer Berlin / Heidelberg.

Book Chapters

- Mateus, D., Wachinger, C., Atasoy, S., Schwarz, L., and Navab, N. (2012). Learning manifolds: design analysis for medical applications. In Susuki, K., editor, *Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis*. IGI Global.

References

[Allain et al., 2009] Allain, B., Hu, M., Lovat, L., Cook, R., Ourselin, S., and Hawkes, D. (2009). Biopsy site re-localisation based on the computation of epipolar lines from two previous endoscopic images. In Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., and Taylor, C., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 5761 of *Lecture Notes in Computer Science*, pages 491–498. Springer Berlin / Heidelberg. 7, 89, 112

[Allain et al., 2010] Allain, B., Hu, M., Lovat, L., Cook, R., Vercauteren, T., Ourselin, S., and Hawkes, D. (2010). A system for biopsy site re-targeting with uncertainty in gastroenterology and oropharyngeal examinations. In Jiang, T., Navab, N., Pluim, J., and Viergever, M., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 6362 of *Lecture Notes in Computer Science*, pages 514–521. Springer Berlin / Heidelberg. 7, 89, 112

[André et al., 2010a] André, B., Vercauteren, T., Buchner, A., Shahid, M., Wallace, M., and Ayache, N. (2010a). An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis. In Jiang, T., Navab, N., Pluim, J., and Viergever, M., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 6362 of *Lecture Notes in Computer Science*, pages 480–487. Springer Berlin / Heidelberg. 7

[André et al., 2010b] André, B., Vercauteren, T., Buchner, A., Wallace, M., and Ayache, N. (2010b). Endomicroscopic Video Retrieval using Mosaicing and Visual Words. *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1419–1422. 7, 14

[André et al., 2011] André, B., Vercauteren, T., Buchner, A., Wallace, M., and Ayache, N. (2011). A smart atlas for endomicroscopy using automated video retrieval. *Medical Image Analysis*, 15(4):460 – 476. 14

[André et al., 2009a] André, B., Vercauteren, T., Perchant, A., Buchner, A., Wallace, M., and Ayache, N. (2009a). Endomicroscopic image retrieval and classification

using invariant visual features. In *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 346–349. IEEE. 7, 14, 102

[André et al., 2009b] André, B., Vercauteren, T., Perchant, A., Buchner, A., Wallace, M., and Ayache, N. (2009b). Introducing space and time in local feature-based endomicroscopic image retrieval. In Caputo, B., Müller, H., Syeda-Mahmood, T., Duncan, J., Wang, F., and Kalpathy-Cramer, J., editors, *Proceedings of Medical Content-based Retrieval for Clinical Decision Support - MICCAI Workshop, (MCBR-CDS)*, volume 5853 of *Lecture Notes in Computer Science*, pages 18–30. Springer Berlin / Heidelberg. 7

[Arnold et al., 2009] Arnold, M., Ghosh, A., Lacey, G., Patchett, S., and Mulcahy, H. (2009). Indistinct frame detection in colonoscopy videos. In *Proceedings of IEEE International Machine Vision and Image Processing Conference (IMVIP)*, pages 47 –52. 13

[Aronszajn and MA., 1950] Aronszajn, N. and MA., H. U. C. (1950). *Theory of reproducing kernels*. Defense Technical Information Center. 36, 38, 39

[Atasoy et al., 2009] Atasoy, S., Glocker, B., Giannarou, S., Mateus, D., Meining, A., Yang, G.-Z., and Navab, N. (2009). Probabilistic region matching in narrow-band endoscopy for targeted optical biopsy. In Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., and Taylor, C., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 5761 of *Lecture Notes in Computer Science*, pages 499–506. Springer Berlin / Heidelberg. 7, 14, 19, 89, 101

[Atasoy et al., 2010a] Atasoy, S., Mateus, D., Georgiou, A., Navab, N., and Yang, G.-Z. (2010a). Wave interference for pattern description. In Kimmel, R., Klette, R., and Sugimoto, A., editors, *Proceedings of Asian Conference on Computer Vision (ACCV)*, volume 6493 of *Lecture Notes in Computer Science*, pages 41–54. Springer Berlin / Heidelberg. 12, 20

[Atasoy et al., 2010b] Atasoy, S., Mateus, D., Lallemand, J., Meining, A., Yang, G.-Z., and Navab, N. (2010b). Endoscopic video manifolds. In Jiang, T., Navab, N., Pluim, J., and Viergever, M., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 6362 of *Lecture Notes in Computer Science*, pages 437–445. Springer Berlin / Heidelberg. 8, 18, 19, 24, 56, 66

[Atasoy et al., 2011] Atasoy, S., Mateus, D., Meining, A., Yang, G.-Z., and Navab, N. (2011). Targeted optical biopsies for surveillance endoscopies. In Fichtinger, G., Martel, A., and Peters, T., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 6893 of *Lecture Notes in Computer Science*, pages 83–90. Springer Berlin / Heidelberg. 8, 18, 19, 24, 56, 88, 98

[Atasoy et al., 2012] Atasoy, S., Mateus, D., Meining, A., Yang, G.-Z., and Navab, N. (2012). Endoscopic video manifolds for targeted optical biopsy. *IEEE Transactions on Medical Imaging*, 31:637–653. 8, 18, 19, 24, 56

[Bajbouj et al., 2010] Bajbouj, M., Delius, S., Becker, V., Jung, A., and Meining, A. (2010). Confocal Laser Scanning Endomicroscopy for in vivo Histopathology of the Gastrointestinal Tract and Beyond–An update. *Arab Journal of Gastroenterology*, 11(4):181–186. 6

[Bashar et al., 2010] Bashar, M., Kitasaka, T., Suenaga, Y., Mekada, Y., and Mori, K. (2010). Automatic Detection of Informative Frames from Wireless Capsule Endoscopy Images. *Medical Image Analysis*, 14(3):449–470. 13

[Bashar et al., 2008] Bashar, M., Mori, K., Suenaga, Y., Kitasaka, T., and Mekada, Y. (2008). Detecting informative frames from wireless capsule endoscopic video using color and texture features. In Metaxas, D., Axel, L., Fichtinger, G., and Székely, G., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 5242 of *Lecture Notes in Computer Science*, pages 603–610. Springer Berlin / Heidelberg. 13

[Belkin, 2003] Belkin, M. (2003). *Problems of Learning Manifolds*. PhD thesis. 44

[Belkin and Niyogi, 2003] Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396. 23, 25, 29, 30, 43, 44, 45, 51, 54, 55, 56, 57, 78, 88

[Belkin and Niyogi, 2007] Belkin, M. and Niyogi, P. (2007). Convergence of laplacian eigenmaps. *Advances in Neural Information Processing Systems*, 19:129. 43, 44

[Belongie et al., 2002] Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 509–522. 119

[Bengio et al., 2004a] Bengio, Y., Delalleau, O., Le Roux, N., Paiement, J.-F., Vincent, P., and Ouimet, M. (2004a). Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10):2197–2219. 43

[Bengio et al., 2004b] Bengio, Y., Paiement, J. F., Vincent, P., Delalleau, O., Le Roux, N., and Ouimet, M. (2004b). Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *Proceedings of Conference on Advances in Neural Information Processing Systems (NIPS)*, page 177. The MIT Press. 28, 43

[Bergman and Schiffer, 1951] Bergman, S. and Schiffer, M. (1951). Kernel functions and conformal mapping. *Compositio Mathematica*, 8:205–249. 39

[Beyer et al., 1999] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin / Heidelberg. 23, 53, 58

[Bileschi and Wolf, 2007] Bileschi, S. and Wolf, L. (2007). Image representations beyond histograms of gradients: The role of gestalt descriptors. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1 –8. 119

[Carman and Welch, 1992] Carman, G. and Welch, L. (1992). Three-dimensional illusory contours and surfaces. *Nature*, 360:585–587. 125

[Chung, 1997] Chung, F. R. K. (1997). *Spectral Graph Theory*. American Mathematical Sociaty. 44

[Coifman and Lafon, 2006] Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30. 23, 25

[Davies and Bouldin, 1979] Davies, D. and Bouldin, D. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):224–227. 79, 81

[Desolneux et al., 2007] Desolneux, A., Moisan, L., and Morel, J. (2007). *From gestalt theory to image analysis: a probabilistic approach*. Springer Verlag. 118

[Dollar et al., 2006] Dollar, P., Tu, Z., and Belongie, S. (2006). Supervised learning of edges and object boundaries. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 2. 118, 119

[Dubuc and Zucker, 2001a] Dubuc, B. and Zucker, S. (2001a). Complexity, confusion, and perceptual grouping. Part I: The curve-like representation. *Journal of Mathematical Imaging and Vision*, 15(1):55–82. 118, 119

[Dubuc and Zucker, 2001b] Dubuc, B. and Zucker, S. (2001b). Complexity, confusion, and perceptual grouping. Part II: Mapping complexity. *International Journal of Computer Vision (IJCV)*, 42(1):83–115. 118, 119

[Egger et al., 2003] Egger, K., Werner, M., Meining, A., Ott, R., Allescher, H., Höfler, H., Classen, M., and Rösch, T. (2003). Biopsy surveillance is still necessary in patients with barrett's oesophagus despite new endoscopic imaging techniques. *British Medical Journal*, 52(1):18. 4

[Elder and Goldberg, 2002] Elder, J. and Goldberg, R. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4):324–353. 119

[Ernst Florens Friedrich Chladni, 1787] Ernst Florens Friedrich Chladni (1787). *Entdeckungen über die Theorie des Klanges*. Leipzig: Weidmanns Erben und Reich. 47

[Ernst Florens Friedrich Chladni, 1802] Ernst Florens Friedrich Chladni (1802). *Die Akustik*. Leipzig: Breitkopf und Härtel. 48

[Etyngier et al., 2007] Etyngier, P., Ségonne, F., and Keriven, R. (2007). Active-contour-based image segmentation using machine learning techniques. In Ayache, N., Ourselin, S., and Maeder, A., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 4791 of *Lecture Notes in Computer Science*, pages 891–899. Springer Berlin / Heidelberg. 24

[Figueiredo and Jain, 2002] Figueiredo, M. and Jain, A. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396. 96

[Fitzgerald, 2004] Fitzgerald, R. (2004). Review article: Barrett's oesophagus and associated adenocarcinoma–a uk perspective. *Alimentary pharmacology & therapeutics*, 20:45–49. 3

[Fukunaga and Olsen, 1971] Fukunaga, K. and Olsen, D. (1971). An Algorithm for Finding Intrinsic Dimensionality of Data. *IEEE Transactions on Computers*, 100(2):176–183. 80

[Geisler et al., 2001] Geisler, W., Perry, J., Super, B., and Gallogly, D. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–724. 118

[Georg et al., 2008] Georg, M., Souvenir, R., Hope, A., and Pless, R. (2008). Simultaneous data volume reconstruction and pose estimation from slice samples. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1 –6. 24

[Gheorghe, 2006] Gheorghe, C. (2006). Narrow-band imaging endoscopy for diagnosis of malignant and premalignant gastrointestinal lesions. *Journal of Gastrointestinal and Liver Diseases*, 15(1):77. 5

[Giannarou et al., 2009] Giannarou, S., Visentini-Scarzanella, M., and Yang, G. (2009). Affine-invariant anisotropic detector for soft tissue tracking in minimally invasive surgery. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 1059–1062. IEEE. 102

[Gono et al., 2004] Gono, K., Obi, T., Yamaguchi, M., Ohyama, N., Machida, H., Sano, Y., Yoshida, S., Hamamoto, Y., and Endo, T. (2004). Appearance of enhanced tissue features in narrow-band endoscopic imaging. *Journal of Biomedical Optics*, 9:568. 5

[Gross et al., 2009] Gross, S., Stehle, T., Behrens, A., Auer, R., Aach, T., Winograd, R., Trautwein, C., and Tischendorf, J. (2009). A Comparison of Blood Vessel Features and Local Binary Patterns for Colorectal Polyp Classification. In *SPIE Medical Imaging - Computer-Aided Diagnosis*. SPIE Vol. 7260, Orlando, USA. 88

[Guo et al., 2003] Guo, C., Zhu, S.-C., and Wu, Y. N. (2003). Towards a mathematical theory of primal sketch and sketchability. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 1228 –1235 vol.2. 118

[Hamm et al., 2010] Hamm, J., Ye, D. H., Verma, R., and Davatzikos, C. (2010). GRAM: A Framework for Geodesic Registration on Anatomical Manifolds. *Medical Image Analysis*, 14(5):633–642. 24

[Hartigan and Wong, 1979] Hartigan, J. and Wong, M. (1979). A k-means Clustering Algorithm. *JR Stat. Soc., Ser. C*, 28:100–108. 70, 78, 97

[Häussinger et al., 2004] Häussinger, K. et al. (2004). Recommendations for quality standards in bronchoscopy. *Pneumologie*, 58:344–356. 93

[He et al., 2005] He, X., Yan, S., Hu, Y., Niyogi, P., and Zhang, H. J. (2005). Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 328–340. 58, 59, 88, 98

[Herbrich, 2002] Herbrich, R. (2002). *Learning kernel classifiers*. MIT Press. 34, 42

[Iakovidis et al., 2006] Iakovidis, D., Maroulis, D., and Karkanis, S. (2006). An Intelligent System for Automatic Detection of Gastrointestinal Adenomas in Video Endoscopy. *Computers in Biology and Medicine*, 36(10):1084–1103. 88

[Iakovidis et al., 2008] Iakovidis, D., Tsevas, S., Maroulis, D., and Polydorou, A. (2008). Unsupervised summarisation of capsule endoscopy video. *Intelligent Systems, 2008. IS'08. 4th International IEEE Conference*, 1:3–15. 88

[Iakovidis et al., 2010] Iakovidis, D., Tsevas, S., and Polydorou, A. (2010). Reduction of Capsule Endoscopy Reading Times by Unsupervised Image Mining. *Computerized Medical Imaging and Graphics*, 34(6):471–478. 13, 88

[Jakobson, D. and Nadirashvili, N. and Toth, 2001] Jakobson, D. and Nadirashvili, N. and Toth, J. (2001). Geometric properties of eigenfunctions. *Russian Mathematical Surveys*, 56. 49

[Joachims, 2002] Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory, and algorithms*. Kluwer. 34, 42

[Kac, 1966] Kac, M. (1966). Can one hear the shape of a drum? *The American Mathematical Monthly*, 73(4):1–23. 49

[Kadoury and Paragios, 2010] Kadoury, S. and Paragios, N. (2010). Nonlinear embedding towards articulated spine shape inference using higher-order mrfs. In Jiang, T., Navab, N., Pluim, J., and Viergever, M., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 6363 of *Lecture Notes in Computer Science*, pages 579–586. Springer Berlin / Heidelberg. 24

[Kaltenbach et al., 2008] Kaltenbach, T., Friedland, S., and Soetikno, R. (2008). A randomised tandem colonoscopy trial of narrow band imaging versus white light examination to compare neoplasia miss rates. *Gut*, 57(10):1406. 6

[Karkanis et al., 2002a] Karkanis, S., Iakovidis, D., Karras, D., and Maroulis, D. (2002a). Detection of Lesions in Endoscopic Video using Textural Descriptors on Wavelet Domain supported by Artificial Neural Network Architectures. In *Proceedings of IEEE International Conference on Image Processing*, volume 2, pages 833–836. 88

[Karkanis et al., 2002b] Karkanis, S., Iakovidis, D., Maroulis, D., Magoulas, G., and Theofanous, N. (2002b). Tumor Recognition in Endoscopic Video Images Using Artificial Neural Network Architectures. In *Proceedings of IEEE Euromicro Conference*, volume 2, pages 423–429. 88

[Koffka, 1999] Koffka, K. (1999). *Principles of Gestalt psychology*. Routledge. 12, 117, 118, 125

[Lanckriet, G. R. G., Cristianini, N., Bartlett, P. L., Ghaoui, L. E., Jordan, 2004] Lanckriet, G. R. G., Cristianini, N., Bartlett, P. L., Ghaoui, L. E., Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, pages 27–72. 34, 42

[Larkin, 1983] Larkin, F. (1983). The weak Gaussian distribution as a means of localization in Hilbert space. *Applied Nonlinear Functional Analysis*, pages 145–177. 39

[Lekadir et al., 2006] Lekadir, K., Elson, D., Requejo-Isidro, J., Dunsby, C., McGinty, J., Galletly, N., Stamp, G., French, P., and Yang, G.-Z. (2006). Tissue characterization using dimensionality reduction and fluorescence imaging. In Larsen, R., Nielsen, M., and Sporring, J., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 4191 of *Lecture Notes in Computer Science*, pages 586–593. Springer Berlin / Heidelberg. 24

[Leordeanu and Hebert, 2005] Leordeanu, M. and Hebert, M. (2005). A spectral technique for correspondence problems using pairwise constraints. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1482–1489. IEEE. 107

[Levine, 1983] Levine, I. (1983). *Quantum Chemistry*. Allyn and Bacon Boston, Massachusetts. 40

[Levy, 2006] Levy, B. (2006). Laplace-Beltrami Eigenfunctions Towards an Algorithm That "Understands" Geometry. *IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06)*, pages 13–13. 47, 49

[Li, 2009] Li, S. (2009). *Markov random field modeling in image analysis*. Springer-Verlag New York Inc. 103

[Lin and Zha, 2008] Lin, T. and Zha, H. (2008). Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):796–809. 24

[Lowe, 1999] Lowe, D. (1999). Object recognition from local scale-invariant features. 2:1150–1157. 102, 118

[Lowe, 2004] Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110. 102, 107, 108, 118

[Maroulis et al., 2003] Maroulis, D., Iakovidis, D., Karkanis, S., and Karras, D. (2003). CoLD: a Versatile Detection System for Colorectal Lesions in Endoscopy Video-frames. *Computer Methods and Programs in Biomedicine*, 70(2):151–166. 88

[Mateus et al., 2012] Mateus, D., Wachinger, C., Atasoy, S., Schwarz, L., and Navab, N. (2012). Learning manifolds: design analysis for medical applications. In Susuki, K., editor, *Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis*. IGI Global. 24

[McDonald, 1965] McDonald, J. E. (1965). Maxwellian interpretation of the Laplacian. *Am. J. Phys*, 33:706–711. 43

[Meining et al., 2010] Meining, A., Atasoy, S., Chung, A., Navab, N., and Yang, G. (2010). "eye-tracking" for assessment of image perception in gastrointestinal endoscopy with narrow-band imaging compared with white-light endoscopy. *Endoscopy*, 42(8):652–655. 6, 12

[Meining et al., 2009] Meining, A., Atasoy, S., Navab, N., Chung, A. J., and Yang, G.-Z. (2009). Targeted optical biopsies for surveillance endoscopies. In *GASTRO*, London, UK. 12

[Meining et al., 2007] Meining, A., Bajbouj, M., Delius, S., and Prinz, C. (2007). Confocal Laser Scanning Microscopy for in Vivo Histopathology of the Gastrointestinal Tract. *Arab Journal of Gastroenterology*, 8(1):1–4. 4, 6

[Mercer, 1909] Mercer, J. (1909). Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 209(441-458):415–446. 36

[Mikolajczyk and Schmid, 2005] Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630. 102, 107

[Mikolajczyk et al., 2005] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Van Gool, L. (2005). A comparison of affine region detectors. *International Journal of Computer Vision (IJCV)*, 65(1/2):43–72. 101

[Moore, 1916] Moore, E. H. (1916). On properly positive Hermitian matrices. *Bull. Amer. Math. Soc*, 23(59):66–67. 38

[Mountney et al., 2009] Mountney, P., Giannarou, S., Elson, D., and Yang, G. (2009). Probabilistic Region Matching in Narrow-Band Endoscopy for Targeted Optical Biopsy. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 5761 of *LNCS*, pages 483–490. Springer. 7, 89, 112

[Murota, 2003] Murota, K. (2003). *Discrete convex analysis*. Number 10. Society for Industrial Mathematics. 107

[Murray, 1988] Murray, J. D. (1988). How the leopard gets its spots. *Scientific American*, 258(3):80—-87. 49

[Murray, 1993] Murray, J. D. (1993). *Mathematical Biology, 2nd Corrected Edition*. Springer. 49

[Nonaka et al., 2006] Nonaka, S., Saito, Y., Gotoda, T., Kozu, T., Matsuda, T., Oda, I., Suzuki, H., and Saito, D. (2006). Narrow band imaging (nbi) system is promising device to detect superficial pharyngeal cancer at an early stage in patients with esophageal squamous cell carcinoma (scc). *Gastrointestinal Endoscopy*, 63(5). 5

[Oh et al., 2007] Oh, J., Hwang, S., Lee, J., Tavanapong, W., Wong, J., and de Groen, P. (2007). Informative Frame Classification for Endoscopy Video. *Medical Image Analysis*, 11(2):110–127. 13

[Paggi et al., 2009] Paggi, S., Radaelli, F., Amato, A., Meucci, G., Mandelli, G., Imperiali, G., Spinzi, G., Terreni, N., Lenoci, N., and Terruzzi, V. (2009). The impact of narrow band imaging in screening colonoscopy: a randomized controlled trial. *Clinical Gastroenterology and Hepatology*, 7(10):1049–1054. 6

[Pearl, 1988] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann. 107

[Peinecke et al., 2007] Peinecke, N., Wolter, F., and Reuter, M. (2007). Laplace spectra as fingerprints for image recognition. *Computer-Aided Design*, 39(6):460–476. 49

[Pless and Souvenir, 2009] Pless, R. and Souvenir, R. (2009). A Survey of Manifold Learning for Images. *IPSJ Transactions on Computer Vision and Applications*, 1(0):83–94. 24

[Rastogi et al., 2009] Rastogi, A., Keighley, J., Singh, V., Callahan, P., Bansal, A., Wani, S., and Sharma, P. (2009). High accuracy of narrow band imaging without magnification for the real-time characterization of polyp histology and its comparison with high-definition white light colonoscopy: a prospective study. *Am J Gastroenterol*, 104(10):2422–30. 6

[Ren et al., 2005] Ren, X., Fowlkes, C., and Malik, J. (2005). Scale-invariant contour completion using conditional random fields. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1214–1221. 118

[Reuter et al., 2006] Reuter, M., Wolter, F., and Peinecke, N. (2006). Laplace-Beltrami spectra as Shape-DNA of surfaces and solids. *Computer-Aided Design*, 38(4):342–366. 49

[Rex and Helbig, 2007] Rex, D. and Helbig, C. (2007). High yields of small and flat adenomas with high-definition colonoscopes using either white light or narrow band imaging. *Gastroenterology*, 133(1):42–47. 6

[Roger and Johnson, 1990] Roger, H. A. and Johnson, C. R. (1990). *Matrix analysis*. Cambridge University Press, Cambridge, UK. 27

[Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323. 23, 25, 29, 30, 43

[Rustamov, 2007] Rustamov, R. M. (2007). Laplace-beltrami eigenfunctions for deformation invariant shape representation. In *Proceedings of the fifth Eurographics symposium on Geometry processing*, pages 225–233, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association. 49

[Scholkopf, B., Tsuda, K., Vert, 2004] Scholkopf, B., Tsuda, K., Vert, J.-P. (2004). *Kernel methods in computational biology*. MIT Press. 34, 42, 43

[Schwartz, 1950] Schwartz, L. (1950). *Theory des distributions*. Paris, vols. 1, 2 edition. 35

[Schwarz et al., 2010] Schwarz, L., Mateus, D., and Navab, N. (2010). Multiple-activity human body tracking in unconstrained environments. In Perales, F. and Fisher, R., editors, *Articulated Motion and Deformable Objects*, volume 6169 of *Lecture Notes in Computer Science*, pages 192–202. Springer Berlin / Heidelberg. 24

[Serre et al., 2007] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411. 118

[Seung and Lee, 2000] Seung, H. and Lee, D. (2000). The manifold ways of perception. *Science*, 290(5500):2268. 23

[Sharma et al., 2006] Sharma, P., Bansal, A., Mathur, S., Wani, S., Cherian, R., McGregor, D., Higbee, A., Hall, S., and Weston, A. (2006). The utility of a novel narrow band imaging endoscopy system in patients with barrett's esophagus. *Gastrointestinal endoscopy*, 64(2):167–175. 3

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905. 54

[Sivic and Zisserman, 2009] Sivic, J. and Zisserman, A. (2009). Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):591–606. 14, 102, 118

[Smale et al., 2010] Smale, S., Rosasco, L., Bouvrie, J., Caponnetto, A., and Poggio, T. (2010). Mathematics of the neural response. *Foundations of Computational Mathematics*, 10(1):67–91. 42

[Souvenir and Pless, 2007] Souvenir, R. and Pless, R. (2007). Image distance functions for manifold learning. *Image and Vision Computing*, 25(3):365–373. 24

[Sparks and Madabhushi, 2010] Sparks, R. and Madabhushi, A. (2010). Novel morphometric based classification via diffeomorphic based shape representation using manifold learning. In Jiang, T., Navab, N., Pluim, J., and Viergever, M., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 6363 of *Lecture Notes in Computer Science*, pages 658–665. Springer Berlin / Heidelberg. 24

[Sparks and Madabhushi, 2011] Sparks, R. and Madabhushi, A. (2011). Out-of-sample extrapolation using semi-supervised manifold learning (ose-ssl): Content-based image retrieval for prostate histology grading. In *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 734–737. IEEE. 24

[Stehle et al., 2009] Stehle, T., Auer, R., Gross, S., Behrens, A., Wulff, J., Aach, T., Winograd, R., Trautwein, C., and Tischendorf, J. (2009). Classification of Colon Polyps in NBI Endoscopy Using Vascularization Features. In *SPIE Medical Imaging - Computer-Aided Diagnosis*. SPIE Vol. 7260, Orlando, USA. 88

[Su et al., 2006] Su, M., Hsu, C., Ho, Y., Chen, P., Lin, C., and Chiu, C. (2006). Comparative study of conventional colonoscopy, chromoendoscopy, and narrow-band imaging systems in differential diagnosis of neoplastic and nonneoplastic colonic polyps. *The American journal of gastroenterology*, 101(12):2711–2716. 6

[Suzuki et al., 2010] Suzuki, K., Zhang, J., and Xu, J. (2010). Massive-Training Artificial Neural Network Coupled with Laplacian-Eigenfunction-Based Dimensionality Reduction for Computer-Aided Detection of Polyps in CT Colonography. *IEEE Transactions on Medical Imaging*, (99):1. 24

[Tamaki et al., 2010] Tamaki, T., Yoshimuta, J., Takeda, T., Raytchev, B., Kaneda, K., Yoshida, S., Takemura, Y., and Tanaka, S. (2010). A system for colorectal tumor classification in magnifying endoscopic nbi images. In Kimmel, R., Klette, R., and Sugimoto, A., editors, *Proceedings of Asian Conference on Computer Vision (ACCV)*, volume 6493 of *Lecture Notes in Computer Science*, pages 452–463. Springer Berlin / Heidelberg. 14, 88

[Tenenbaum et al., 2000] Tenenbaum, J. B., Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319. 23, 25, 28, 30, 43, 55

[Tischendorf et al., 2007] Tischendorf, J., Wasmuth, H., Koch, A., Hecker, H., Trautwein, C., and Winograd, R. (2007). Value of magnifying chromoendoscopy and

narrow band imaging (nbi) in classifying colorectal polyps: a prospective controlled study. *Endoscopy*, 39(12):1092–1096. 6

[Tiwari et al., 2008] Tiwari, P., Rosen, M., and Madabhushi, A. (2008). Consensus-Locally Linear Embedding (C-LLE): Application to Prostate Cancer Detection on Magnetic Resonance Spectroscopy. In Metaxas, D., Axel, L., Fichtinger, G., and Székely, G., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science, pages 330–338. Springer. 24

[Tiwari et al., 2009] Tiwari, P., Rosen, M., Reed, G., Kurhanewicz, J., and Madabhushi, A. (2009). Spectral Embedding Based Probabilistic Boosting Tree (ScEPTre): Classifying High Dimensional Heterogeneous Biomedical Data. In Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., and Taylor, C., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science, pages 844–851. Springer Berlin / Heidelberg. 24

[Torresani et al., 2008] Torresani, L., Kolmogorov, V., and Rother, C. (2008). Feature correspondence via graph matching: Models and global optimization. In *Proceedings of IEEE European Conference on Computer Vision (ICCV)*, pages 596–609. Springer. 107

[Vallet and Lévy, 2008] Vallet, B. and Lévy, B. (2008). Spectral Geometry Processing with Manifold Harmonics. *Computer Graphics Forum*, 27(2):251–260. 50

[van der Maaten et al., 2009] van der Maaten, L., ostma, E. P., and van den Herik, H. . (2009). Dimensionality reduction: A comparative review. Technical report, TiCC-TR 2009-005, Tilburg University. 24

[Von Luxburg, 2007] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416. 70, 80

[Wachinger and Navab, 2010] Wachinger, C. and Navab, N. (2010). Structural image representation for image registration. *CVPR Workshops, IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 23–30. 24

[Wachinger et al., 2010] Wachinger, C., Yigitsoy, M., and Navab, N. (2010). Manifold learning for image-based breathing gating with application to 4d ultrasound. In Jiang, T., Navab, N., Pluim, J., and Viergever, M., editors, *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 6362 of *Lecture Notes in Computer Science*, pages 26–33. Springer Berlin / Heidelberg. 23

[Wani and Sharma, 2006] Wani, S. and Sharma, P. (2006). The rationale for screening and surveillance of barrett's metaplasia. *Best Practice & Research Clinical Gastroenterology*, 20(5):829–842. 3

[Watt and Phillips, 2000] Watt, R. and Phillips, W. (2000). The function of dynamic grouping in vision. *Trends in Cognitive Sciences*, 4(12):447–454. 119

[Wibisono et al., 2010] Wibisono, A., Bouvrie, J., Rosasco, L., and Poggio, T. (2010). Learning and invariance in a family of hierarchical kernels. Technical report, MIT-CSAIL-TR-2010-035/CBCL-290, Massachusetts Institute of Technology, Cambridge, MA. 42

[Williams and Jacobs, 1997] Williams, L. and Jacobs, D. (1997). Stochastic completion fields: A neural model of illusory contour shape and salience. *Neural Computation*, 9(4):837–858. 118, 119

[Zass and Shashua, 2008] Zass, R. and Shashua, A. (2008). Probabilistic graph and hypergraph matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE. 107, 108

[Zhang et al., 2006] Zhang, Q., Souvenir, R., and Pless, R. (2006). On manifold structure of cardiac mri data: Application to segmentation. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1092–1098. IEEE. 24