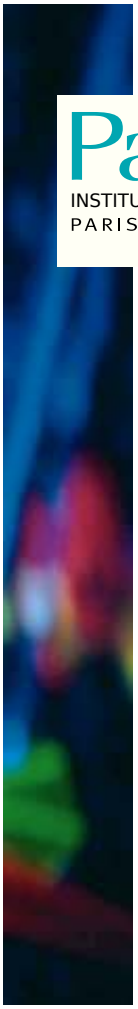
Smart atlas for endomicroscopy diagnosis support: a clinical application of content-based image retrieval

Barbara André



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale n°84:
Sciences et technologies de l'information et de la communication

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité
"Informatique temps-réel, robotique et automatique"

présentée et soutenue publiquement par

Barbara ANDRE

le 12 Octobre 2011

Smart Atlas for Endomicroscopy Diagnosis Support:

A Clinical Application of Content-Based Image Retrieval

Directeur de thèse: **Nicholas AYACHE**
Co-encadrement de la thèse: **Tom VERCAUTEREN**

Jury

Josiane ZERUBIA, Ariana Research Team, INRIA Sophia Antipolis	Présidente
Antonio CRIMINISI, Microsoft Research Cambridge	Rapporteur
Sébastien OURSELIN, University College London	Rapporteur
Michael B. WALLACE, Mayo Clinic, Jacksonville	Examinateur
Christian BARILLOT, VisAGeS Research Team, IRISA Rennes	Examinateur
Nicholas AYACHE, Asclepios Research Team, INRIA Sophia Antipolis	Directeur
Tom VERCAUTEREN, Mauna Kea Technolgies, Paris	Examinateur

T H È S E

MINES ParisTech
Centre de Mathématiques Appliquées
Rue Claude Daunesse B.P. 207, 06904 Sophia Antipolis Cedex, France

Remerciements

Je tiens tout d'abord à remercier chaleureusement Nicholas Ayache et Tom Vercauteren pour m'avoir encadrée et guidée tout au long de ma thèse avec tant d'attention et d'intelligence; j'ai particulièrement apprécié chez Nicholas son enthousiasme communicatif et son instinct visionnaire, et chez Tom la justesse de ses conseils, son aide précieuse dans de nombreux domaines et sa grande disponibilité. Je suis également très reconnaissante à Sacha Loiseau de m'avoir accueillie à Mauna Kea Technologies, son soutien fut indispensable à l'aboutissement de cette thèse.

Je remercie sincèrement tous les membres de mon jury de thèse, en particulier Josiane Zerubia, Antonio Criminisi et Sébastien Ourselin pour avoir pris le temps de lire ce manuscrit avec attention et de me transmettre leurs retours très constructifs. Merci à Christian Barillot pour avoir accepté de faire partie de mon jury de thèse. Je tiens à exprimer ma profonde gratitude à Michael Wallace qui s'est rendu disponible pour apporter au sein du jury un éclairage complémentaire sous l'angle de son expertise médicale.

Dans la société Mauna Kea Technologies, j'ai eu plaisir à travailler avec différentes personnes qui m'ont beaucoup appris. Aymeric Perchant, qui a participé à l'encadrement du début de ma thèse, m'a toujours fait part de remarques très pertinentes sur ce sujet de recherche. Je remercie par ailleurs tous les membres de l'équipe R&D pour le soutien et l'aide qu'ils m'ont apportés, notamment Guillaume Schmid, Nicolas Savoire et François Lacombe. Merci à Benoît Mariaux pour m'avoir aidée à mettre en ligne l'outil VSS. Dans l'équipe des Affaires Cliniques, je tiens à remercier Anne Osdoit, France Schwarz, Céline Peltier, Cindy Warren et Cécile Redon pour leur aide considérable dans la construction des bases de données, l'interprétation des données cliniques et la communication régulière avec les médecins.

Au sein de l'équipe Asclepios à l'INRIA de Sophia Antipolis, j'ai beaucoup apprécié les échanges prolifiques que j'ai pu avoir avec les chercheurs et les ingénieurs, en particulier avec Grégoire Malandain, Xavier Pennec, Ezequiel Geremia, Erik Pernod, Liliane Ramus, Tommaso Mansi et Florence Billet. Je remercie d'autre part Isabelle Strobant pour sa gentillesse et pour avoir facilité l'organisation de mes différents voyages entre Paris et Sophia Antipolis.

Je suis profondément reconnaissante à Michael Wallace et Anna Buchner, qui ont contribué à l'acquisition et l'annotation de la majorité des données cliniques utilisées dans cette thèse. Ils se sont toujours rendus disponibles pour faire avancer la recherche et apporter leur expertise médicale sur ce sujet. Merci à Waseem Shahid pour m'avoir accueillie à la Mayo Clinic pendant mes séjours en Floride. Je tiens également à remercier tous les médecins qui ont contribué à la construction de la vérité terrain sur la similarité perçue en utilisant l'outil VSS, merci particulièrement

à Vani Konda, Waseem Shahid et Emmanuel Coron.

Enfin, je remercie avec émotion les personnes qui me sont proches et qui m'ont toujours soutenue, ma famille, Julien.

Table of Contents

Chapter 1

Introduction

Table of Contents

Foreword This thesis stems from a CIFRE agreement[1] with Asclepios research team at INRIA Sophia Antipolis, `http://www-sop.inria.fr/asclepios`, and the company Mauna Kea Technologies, Paris, `http://www.maunakeatech.com`, which is specialized in the development of *in vivo* cellular imaging systems for biomedical and medical applications.

French summary

L'Endomicroscopie Confocale par Minisondes (ECM) est une technologie récente qui permet l'observation dynamique des tissus au niveau cellulaire, in vivo et in situ*, pendant une endoscopie. Grâce à ce nouveau système d'imagerie, les médecins endoscopistes ont la possibilité de réaliser des "biopsies optiques" non invasives. Les biopsies traditionnelles impliquent le diagnostic* ex vivo *d'images histologiques par des médecins pathologistes. Le diagnostic* in vivo *d'images ECM est donc un véritable challenge pour les endoscopistes, qui ont en général seulement un peu d'expertise en anatomopathologie. Les images ECM sont néammoins de nouvelles images, qui ressemblent visuellement aux images histologiques. Cette thèse a pour but principal d'assister les endoscopistes dans l'interprétation* in vivo *des séquences d'images ECM, en mettant à leur disposition un système de reconnaissance de vidéos endomicroscopiques. Nous proposons de construire un atlas intelligent, capable d'extraire automatiquement dans une base de données, plusieurs vidéos ECM qui ont une apparence similaire à la vidéo requête, mais qui ont déjà été annotées avec différentes métadonnées telles que par exemple le diagnostic histologique.*

[1] CIFRE (Convention Industrielle de Formation par la Recherche / Industrial Agreement for Training via Research) agreements aim at fostering innovative processes and technology transfer between public research organizations and industry by supporting a young researcher based in industry, to complete the PhD. They are administered by ANRT (Association Nationale de la Recherche Technique / National Association for Technical Research), `http://www.anrt.asso.fr`.

Figure 1.1: Left: pCLE miniprobe inserted through the working channel of a standard endoscope. **Right:** Setup of pCLE imaging system (Cellvizio, Mauna Kea Technologies).

1.1 A Smart Atlas for Endomicroscopy: How to Support *In Vivo* Diagnosis of Gastrointestinal Cancers?

In the last decade, the visualization of epithelial tissues at cellular level has been made possible in the living organism by the use of fibered confocal microscopy. In particular, probe-based Confocal Laser Endomicroscopy (pCLE) enables the *in vivo* microscopic imaging of the epithelium during ongoing endoscopy, at real-time frame-rate, and *in situ*, i.e. in contact with the region of interest. The pCLE imaging system is illustrated in Fig. 1.1: a confocal miniprobe, made of tens of thousands of optical fibers, is inserted through the working channel of a standard endoscope to image an optical plane at a fixed distance below the surface of the tissue. The pCLE miniprobe is connected to a proximal laser scanning unit which uses two mirrors to emit, along each optical fiber, an excitation light that is locally absorbed by fluorophores in the tissue. The light which is then emitted by the fluorophores at a longer wavelength is transferred back along the same fiber to a mono-pixel photodetector. As a result, pCLE images with field-of-view ranging from 200 to 600 μm are acquired at a rate of 9 to 18 frames per second, composing image sequences called pCLE videos.

For the endoscopists, the pCLE imaging system is a revolutionary tool which gives them the opportunity to perform non-invasive "optical biopsies", and thus

to establish *in vivo* diagnosis of epithelial cancers. Using pCLE, endoscopists are provided *online* with new images that visually look like the histological images, as shown in Fig. 1.2. Histological images are usually diagnosed *offline* by pathologists. Today, histology remains the "gold standard" for cancer diagnosis. However, *ex vivo* histological diagnosis implies invasive procedures that are potentially dangerous for the patient, and a large proportions of unnecessary biopsies associated with a significant cost. In [Wang 07], Wang and Camilleri pointed out that pCLE enables combined diagnosis and treatment during the endoscopy procedure:

> *The ultimate goal should be that the gastroenterologist-endoscopist be in the driver's seat in the management of patients presenting with mucosal lesions that are appraised thoroughly with endoscopic procedures including histologic characterization, assessment of depth of invasion and surrounding tissues and lymph nodes, and, ultimately, resection in toto using submucosal dissection if necessary.* (p. 1260)

Thus, the *in vivo* diagnosis of epithelial cancers is a challenge for the endoscopists. In particular, the early diagnosis of gastrointestinal cancers, that are a leading cause of cancer death worldwide, is one of the critical challenges. Currently, pCLE is relatively new to many physicians, who are still in the process of defining a taxonomy of the pathologies seen in pCLE images: for a given pathological class, there is a high variability in the appearance of pCLE images. The main goal of this thesis is to assist the endoscopists in the interpretation of pCLE image sequences.

The subjective experience of understanding the pathologies observed in pCLE images would undoubtedly benefit from an objective tool that provides clinically interpretable information to guide the interpretation. When establishing a pathological diagnosis from a new image, physicians typically use similarity-based reasoning: they implicitly rely on *visually* similar cases they have seen in the past. This is the reason why we propose to investigate, for diagnosis support, content-based image retrieval (CBIR) approaches that manipulate low-level visual features. Following the query-by-example model, we aim at developing a retrieval system which automatically extracts, from a *training database*, several pCLE videos that are visually *similar* to the pCLE video of interest, but that have been previously annotated with metadata. A suitable *training database* for pCLE retrieval would contain a sufficiently large number of representative pCLE videos together with their attached metadata, including pathological diagnosis. Such a pCLE retrieval system would thus act like a "Smart Atlas" that opens for the endoscopist a comprehensive book of already diagnosed and annotated pCLE cases at the right pages. The proposed solution is not only an "Atlas" defined by the annotated database, but a "Smart Atlas" able to extract the relevant information. By guiding the endoscopists in making informed decisions during the procedure, the retrieval system would help them in establishing more accurate pCLE diagnoses. The concept of a "Smart Atlas" for pCLE is illustrated in Fig. 1.3.

Because establishing a pCLE diagnosis is an everyday practice for the endoscopist, the retrieval tools may also be used as a training tool that assists the

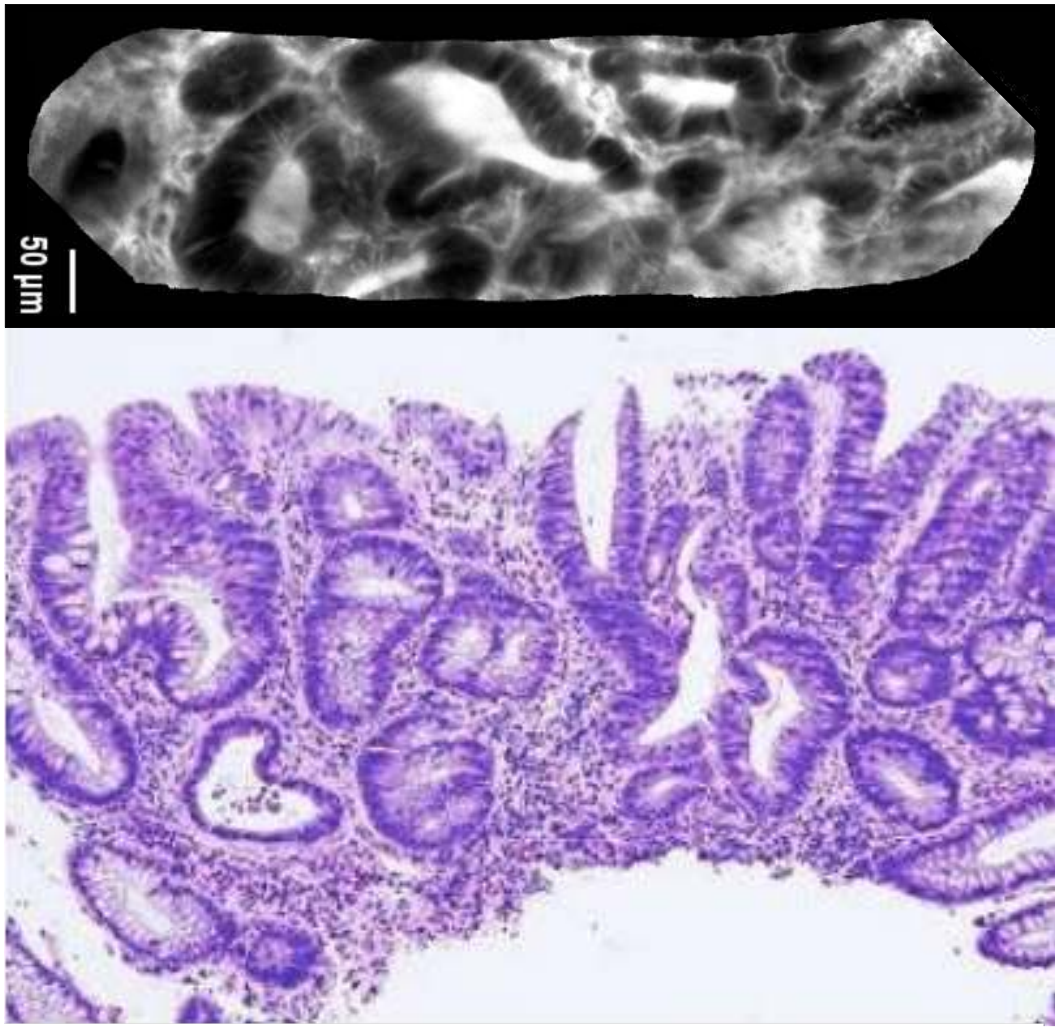


Figure 1.2: Top: pCLE mosaic image obtained from the "optical biopsy" of a suspicious polyp in the colon. **Bottom:** Corresponding histological image obtained from the real biopsy of the same suspicious polyp. The suspicious polyp was diagnosed as tubular adenoma (i.e. malignant). Whereas histological cuts provides images in a transverse plane, pCLE produces images in an "en-face" plane (i.e. a frontal view).

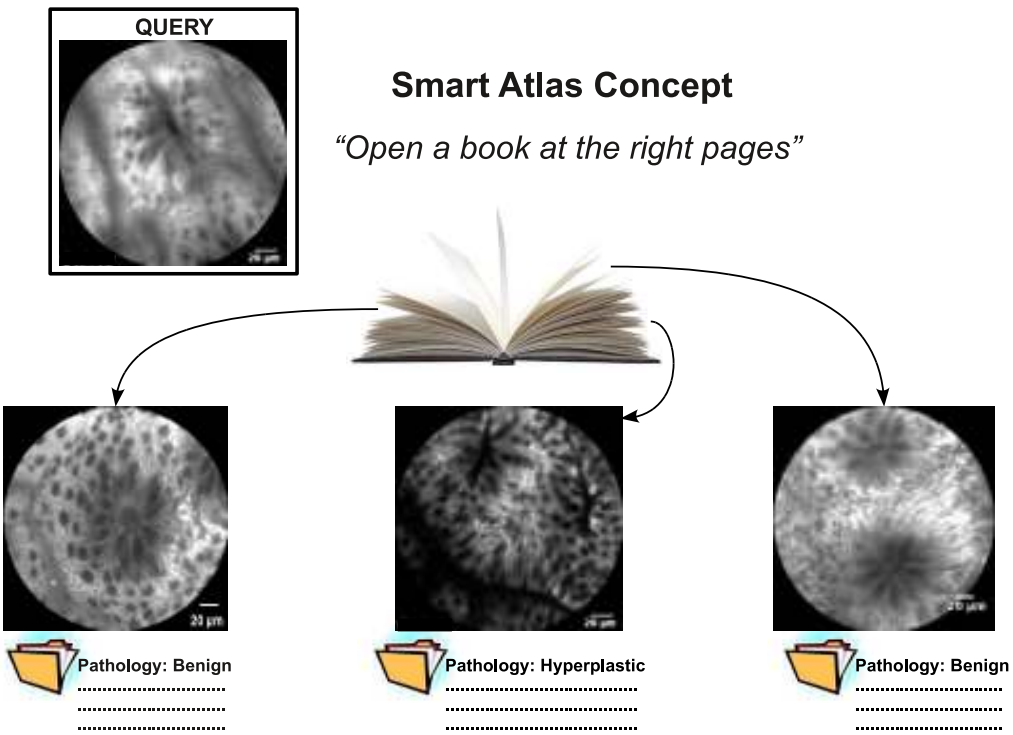


Figure 1.3: Schematic example illustrating the "Smart Atlas" concept for pCLE. The example images are typical pCLE images of colonic polyps. Three annotated pCLE images, visually similar to the query image, are automatically extracted from an annotated *training database* represented by the "Smart Atlas" book. These extracted images are annotated with metadata, such as their pathological diagnosis.

endoscopists in shortening their learning curve. For example, a difficulty level for the interpretation of pCLE videos could be estimated to complement the retrieval outputs. Furthermore, high-level clinical knowledge, such as the visual similarity perceived between pCLE videos or the semantic concepts used to describe pCLE videos, could be included to define an adequate similarity distance between pCLE videos, where *similarity* is thought in terms of visual content *and* semantic annotations. On the relevance of the learned similarity distance would depend the navigation of the endoscopist through the multimodal *training database*. A relevant pCLE retrieval system would also require the *training database* to be sufficiently representative of the variability in the appearance of pCLE cases. Besides, the *training database* could be enriched over time by the endoscopists who may add or annotate new pCLE cases that were never seen before. This should support not only knowledge sharing between the endoscopists, but also pCLE knowledge discovery.

1.2 From Computer Vision to Medical Applications

The CBIR techniques, inherited from the computer vision field, were initially applied to non-medical image databases and for classification purpose. A large review of the state of the art in CBIR is presented by Smeulders et al. [Smeulders 00], who pointed out the need for large representative databases, the problem of retrieval evaluation, and the existence of the semantic gap.

In computer vision, many CBIR methods have been developed. Some are more object-oriented, some are texture-oriented. Boureau et al. [Boureau 10] identified two important phases in CBIR: the coding phase which decomposes the original image features on a dictionary according to desirable properties, e.g. invariance, compactness, sparseness, statistical independence, and the pooling phase which summarizes the resulting codes into a single image signature. Using two annotated databases of natural scenes and objects, they provided a comprehensive evaluation of several combinations of the coding modules, e.g. hard and soft vector quantization or sparse coding, and the pooling schemes, e.g. average-based or maximum-based pooling. In particular, the Bag-of-Visual-Words (BoW) method proposed by Sivic and Zisserman [Sivic 06] is a CBIR method that uses a visual vocabulary, based on vector quantized viewpoint invariant descriptors. The BoW method is relevant for texture retrieval. Indeed, Zhang et al. [Zhang 07] achieved excellent classification results by applying the BoW method to a database of natural texture images. Lazebnik et al. [Lazebnik 06] extended the BoW model to a spatial pyramid matching model that manipulates histograms of image features over image subregions. By extracting both spectral features using the "gist" descriptor of Oliva and Torralba [Oliva 01], and gradient features using the SIFT descriptor of Lowe [Lowe 04], the spatial pyramid matching method achieves high classification accuracy on a large annotated database of fifteen natural scene categories. Proposing another design, Chehade et al. [Chehade 09] used Haralick features descriptors [Haralick 79] for the texture classification of vegetation types in aerial color infra-red images.

Various medical applications of CBIR were proposed in the literature. For example, the Neurobase project, proposed by Barillot et al. [Barillot 04], aims at building an information system that would allow multimodal similarity search in neuroimaging. Müller et al. [Müller 08] presented a benchmark for the evaluation of multimodal CBIR methods on medical databases, according to the "ImageCLEF" medical image retrieval task that includes heterogeneous medical images, from radiography and electrocardiograms to histopathology. The application of CBIR to uterine cervix images was investigated by Greenspan [Greenspan 09] in order to facilitate training and research on uterine cancers. More recently, Simonyan et al. [Simonyan 11] proposed a visual search framework using regions of interest for the immediate retrieval of medical images and the simultaneous localization of anatomical structures. The choice of the appropriate CBIR method highly depends on the targeted medical application. Because discriminative information is dense in the pCLE images, which have a similar appearance to texture images, we will

explore in this thesis a dense version of the BoW model in order to achieve pCLE retrieval.

1.3 Manuscript Organization and Contributions

The present thesis is organized along our published and submitted studies, on which it is largely based. The resulting manuscript progresses from the development of an objective tool for diagnosis support to the learning of higher-level clinical knowledge for training support.

Chapter 2, based on [André 11e], focuses on the main methodological contributions that adjust the standard BoW method for the retrieval of pCLE videos. Built from our submitted clinical article [André 11a], Chapter 3 presents the clinical application of pCLE video retrieval on colonic polyps. This study was presented in a clinical abstract [André 10b] copied in Appendix B. Chapter 4, based on [André 10a], proposes an automated estimation of the difficulty to interpret pCLE videos, from the retrieval results obtained on two different pCLE video databases, the *Colonic Polyps* and the Barrett's Esophagus. This study was presented in a second clinical abstract [André 11b] shown in Appendix C. Finally, in Chapter 5 based on our submitted article [André 11c], more clinical knowledge is included in order to learn semantic and visual similarity between pCLE videos.

We start in Chapter 2 by analyzing the image properties of pCLE videos. Observing that epithelial tissues are characterized by the regularity of the cellular and vascular architectures, we aim at retrieving discriminative texture information coupled with shape information by applying local operators on pCLE images. To serve that purpose, we revisit the BoW method which has been successfully used in many applications of computer vision. The standard BoW method consists of detecting salient image regions from which continuous features are extracted and discretized into "visual words". In order to capture all the discriminative information which is densely distributed in pCLE images, we propose a dense BoW description of these images. Further methodological contributions, using multi-scale description, visual word weighting and the co-occurrence matrix of visual words, are then investigated. Knowing that the images composing a pCLE video are mostly related by viewpoint changes, we leverage a video-mosaicing technique to build a single visual word signature per video. Because of the subjective appreciation of visual similarities between images, it is difficult to have a ground truth for CBIR. If no *visual similarity ground truth* is available, an objective method to evaluate retrieval performance is nearest neighbor classification. We thus evaluate the resulting pCLE retrieval method in an indirect manner, using first a binary pathological classification and then a finer multi-class pathological classification. To avoid bias, leave-one-patient-out cross-validations were performed, where the pathological class of each video is its pCLE diagnosis confirmed by histology. These *indirect retrieval evaluations* show that, on the pCLE video database of colonic polyp, our retrieval method outperforms with statistical significance several state-of-the-art methods in CBIR.

Chapter 3 presents a clinical application of the methodology described in Chapter 2, that specifically addresses the binary classification between malignant and non-malignant colonic polyps. The proposed CBIR-based classification is applied to an extended pCLE video database of colonic polyps, that also contains the videos for which the pCLE diagnosis was in contradiction with the histological diagnosis. Histology was used as gold standard for the differentiation between neoplastic (i.e. malignant) and non-neoplastic (i.e. non-malignant) polyps. We demonstrate that, in terms of binary pathological classification, the performance of the pCLE retrieval system is rather high (accuracy 89.6%, sensitivity 92.5%, specificity 83.3%) and equivalent, with statistical significance, to the offline diagnosis performance of two human expert endoscopists. This chapter also provides a deeper insight into the clinical procedures, from pCLE acquisition protocol and pCLE examination to histological examination.

As pCLE diagnosing is a challenging everyday practice that benefits from experience, our objective in Chapter 4 is to help the endoscopists in shortening their learning curve in pCLE diagnosis. We propose a method to estimate, based on the retrieval results, the difficulty to interpret a pCLE video. Such an estimator could thus be used in a structured training simulator that features difficulty level selection. As a first step toward clinical evaluation, we show that there is a significant relationship between the estimated difficulty and the diagnosis difficulty which has been experienced by multiple endoscopists.

Because our pCLE retrieval method provides visual word signatures that adequately represent pCLE videos, the use of a standard distance on these *visual* signatures already provides relevant results. However, little clinical knowledge has been included to obtain these results. This is the reason why we investigate in Chapter 5 how the incorporation of prior clinical information could enable the learning of the visual similarity distance *and* of pCLE semantics. For the generation of *a visual similarity ground truth*, we develop an online survey tool that allows multiple observers, who are experts in pCLE, to qualitatively estimate the visual similarity that they perceive between pCLE videos. From the perceived similarity data, we are able to learn an adjusted visual similarity distance which we prove to be better than the original retrieval distance. We also use this sparse *visual similarity ground truth* to define "sparse recall" curves and perform *direct retrieval evaluations*, the results of which confirm our first results from *indirect retrieval evaluations*. Finally, in order to learn pCLE semantics, we leverage semantic information from multiple concepts used by the endoscopists to describe pCLE videos. In a first attempt to bridge the semantic gap, we build visual-word-based *semantic* signatures which extract, from low-level visual features, a higher-level clinical knowledge that is directly interpretable by the endoscopist and consistent with respect to perceived similarity.

1.4 List of Publications

This thesis is largely based on the following publications and submitted articles:

Methodological publications (peer-reviewed full papers)

[**André 11c**] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *Learning semantic and visual similarity for endomicroscopy video retrieval*. 2011. Article in submission

[**André 11d**] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *Retrieval evaluation and distance learning from perceived similarity between endomicroscopy videos*. In Proceedings of the 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'11), pages 289–296, 2011

[**André 11e**] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *A smart atlas for endomicroscopy using automated video retrieval*. Medical Image Analysis, volume 15, number 4, pages 460–476, August 2011

[**André 10a**] B. André, T. Vercauteren, A. M. Buchner, M. W. Shahid, M. B. Wallace and N. Ayache. *An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis*. In Proceedings of the 13th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'10), pages 480–487, 2010

[**André 10c**] B. André, T. Vercauteren, M. B. Wallace, A. M. Buchner and N. Ayache. *Endomicroscopic video retrieval using mosaicing and visual words*. In Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'10), pages 1419–1422, 2010

[**André 09b**] B. André, T. Vercauteren, A. Perchant, M. B. Wallace, A. M. Buchner and N. Ayache. *Introducing space and time in local feature-based endomicroscopic image retrieval*. In Proceedings of the MICCAI 2009 Workshop - Medical Content-based Retrieval for Clinical Decision (MCBR-CDS'09), pages 18–30, 2009

[**André 09a**] B. André, T. Vercauteren, A. Perchant, M. B. Wallace, A. M. Buchner and N. Ayache. *Endomicroscopic image retrieval and classification using invariant visual features*. In Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'09), pages 346–349, 2009

Clinical publications (peer-reviewed full papers)

[**André 11a**] B. André, T. Vercauteren, A. M. Buchner, M. Krishna, N. Ayache and M. B. Wallace. *Video retrieval software for automated classification of probe-based confocal laser endomicroscopy on colorectal polyps*. 2011. Article in submission

Clinical publications (abstracts)

[**André 11b**] B. André, T. Vercauteren, A. M. Buchner, M. W. Shahid, M. B. Wallace and N. Ayache. *Toward a structured training system for probe-based confocal laser endomicroscopy (pCLE) on Barrett's esophagus: a video retrieval approach to estimate diagnosis difficulty.* Gastrointestinal Endosccopy (DDW 2011), volume 73, number 4 Suppl, page AB398, 2011. Selected Video Abstract available at `http://www.youtube.com/watch?v=RVy-0Bxx9EQ`

[**André 10b**] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *Endomicroscopic video retrieval approach to support diagnostic differentiation between neoplastic and non-neoplastic colonic polyps.* Gastroenterology (DDW 2010), volume 138, number 5 Suppl, pages S–514, May 2010

Chapter 2

Adjusting Bag-of-Visual-Words for Endomicroscopy Video Retrieval

Table of Contents

Based on: [André 11e] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *A smart atlas for endomicroscopy using automated video retrieval.* Medical Image Analysis, volume 15, number 4, pages 460–476, August 2011. **Additional material available in** [André 09a], [André 09b] and [André 10a].

To support the challenging task of early epithelial cancer diagnosis from in vivo *endomicroscopy, we propose a content-based video retrieval method that uses an expert-annotated database. Motivated by the recent successes of non-medical content-based image retrieval, we first adjust the standard Bag-of-Visual-Words method to handle single endomicroscopic images. A local dense multi-scale description is proposed to keep the proper level of invariance, in our case to translations, in-plane rotations and affine transformations of the intensities. Since single images may have an insufficient field of view to make a robust diagnosis, we introduce a video-mosaicing technique that provides large field-of-view mosaic images from input video sequences. To remove potential outliers, retrieval is followed by a geometrical approach that captures a statistical description of the spatial relationships between the local features. Building on image retrieval, we then focus on efficient video retrieval. Our approach avoids the main time-consuming parts of the video-mosaicing by relying on coarse registration results only to account for spatial overlap between images taken at different times. To evaluate the retrieval, we perform a simple nearest neighbor classification with leave-one-patient-out cross-validation. From the results of binary and multi-class classification, we show that our approach outperforms, with statistical significance, several state-of-the art methods. We obtain a binary classification accuracy of* 94.2 *%, which is quite close to clinical expectations.*

French summary

Afin d'aider les endoscopistes dans le diagnostic des cancers précoces de l'épithélium à partir de l'endomicroscopie in vivo*, nous proposons une méthode de reconnaissance de vidéos par le contenu qui s'appuie sur une base de données annotée par des experts. Motivés par les succès récents de la reconnaissance d'images par le contenu dans le domaine non médical, nous commençons par ajuster la méthode des Sacs de Mots Visuels pour la reconnaissance d'images endomicroscopiques isolées. Une description dense multi-échelle est proposée pour assurer le niveau d'invariance recherché, à savoir dans notre cas l'invariance par translation, par rotation dans le plan et par transformation affine des intensités. Etant donné que le champ de vue des images isolées peut s'avérer insuffisant pour l'établissement d'un diagnostic robuste, nous introduisons une technique de mosaïcage produisant des images de grand champ à partir de séquences vidéos. Afin d'éliminer d'éventuels intrus, l'étape de reconnaissance est suivie par une étape de vérification géométrique qui utilise une description statistique des relations spatiales entre les mots visuels. La reconnaissance d'images isolées étant résolue, nous visons alors une méthode efficace pour la reconnaissance de vidéos. Notre approche contourne le problème du temps de calcul relativement long du mosaïcage en ne prenant en compte que les résultats de translation, obtenus en temps réel, pour calculer le recouvrement spatial entre les images prises à différents instants. A partir des résultats de classification binaire et multi-classe, nous montrons que notre approche surpasse de manière significative*

plusieurs méthodes de l'état de l'art. Pour la classification binaire, nous obtenons une précision de 94.2*%, ce qui est très proche des exigences cliniques.*

2.1 Introduction

Standard endoscopic imaging allows only the diagnosis of disease states with moderate levels of certainty, as pointed out by Norfleet et al. [Norfleet 88, Rastogi 09]. Consequently, biopsies are frequently performed during endoscopy procedures in order to establish, *ex vivo*, a definitive diagnosis. However, biopsies are invasive procedures which may be unnecessary, as some resected specimens are ultimately found to be normal tissue. Furthermore, the need for confirmatory biopsy delays the diagnosis and often requires a separate endoscopic procedure to be performed for treatment.

With the recent technology of probe-based confocal laser endomicroscpy (pCLE), physicians are able to image the epithelium at microscopic level and in real time during an ongoing endoscopy procedure. As mentioned by Wallace and Fockens [Wallace 09], the main task for the endoscopists is to establish a diagnosis *in vivo* from the acquired pCLE videos, by relating a given appearance of the epithelium to a specific pathology.

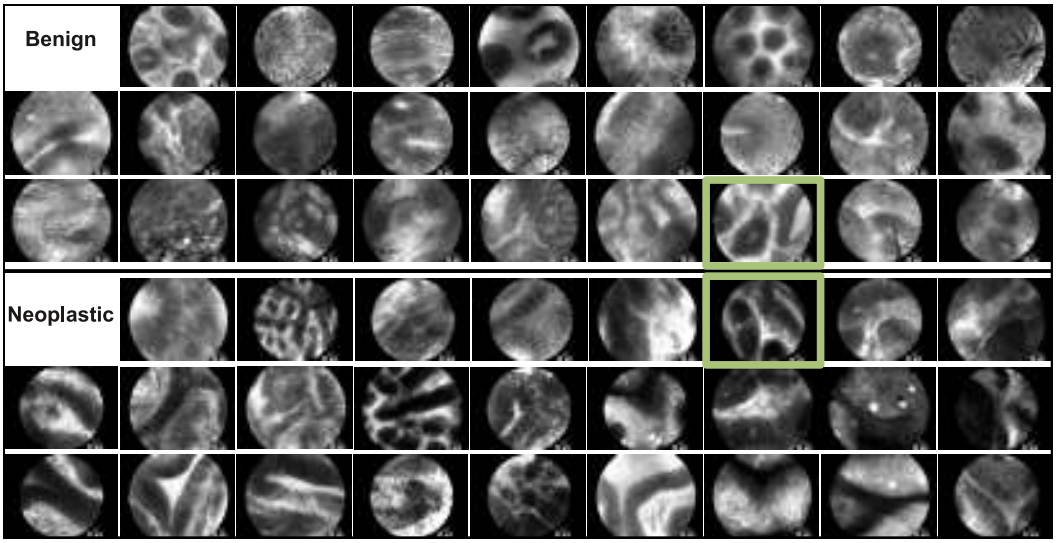


Figure 2.1: pCLE image samples from our database of colonic polyps. The pCLE images have a diameter of approximately 500 pixels that corresponds to a field of view of 240 μm. Images of the polyps diagnosed as benign are on the top, whereas those diagnosed as neoplastic are on the bottom. The closer to the boundary the images are, the less obvious is their diagnosis according to their visual appearance. In particular, the two framed images might look similar although they belong to different pathological classes. This panel also illustrates the large intra-class variability, within the benign class as well as within the neoplastic class.

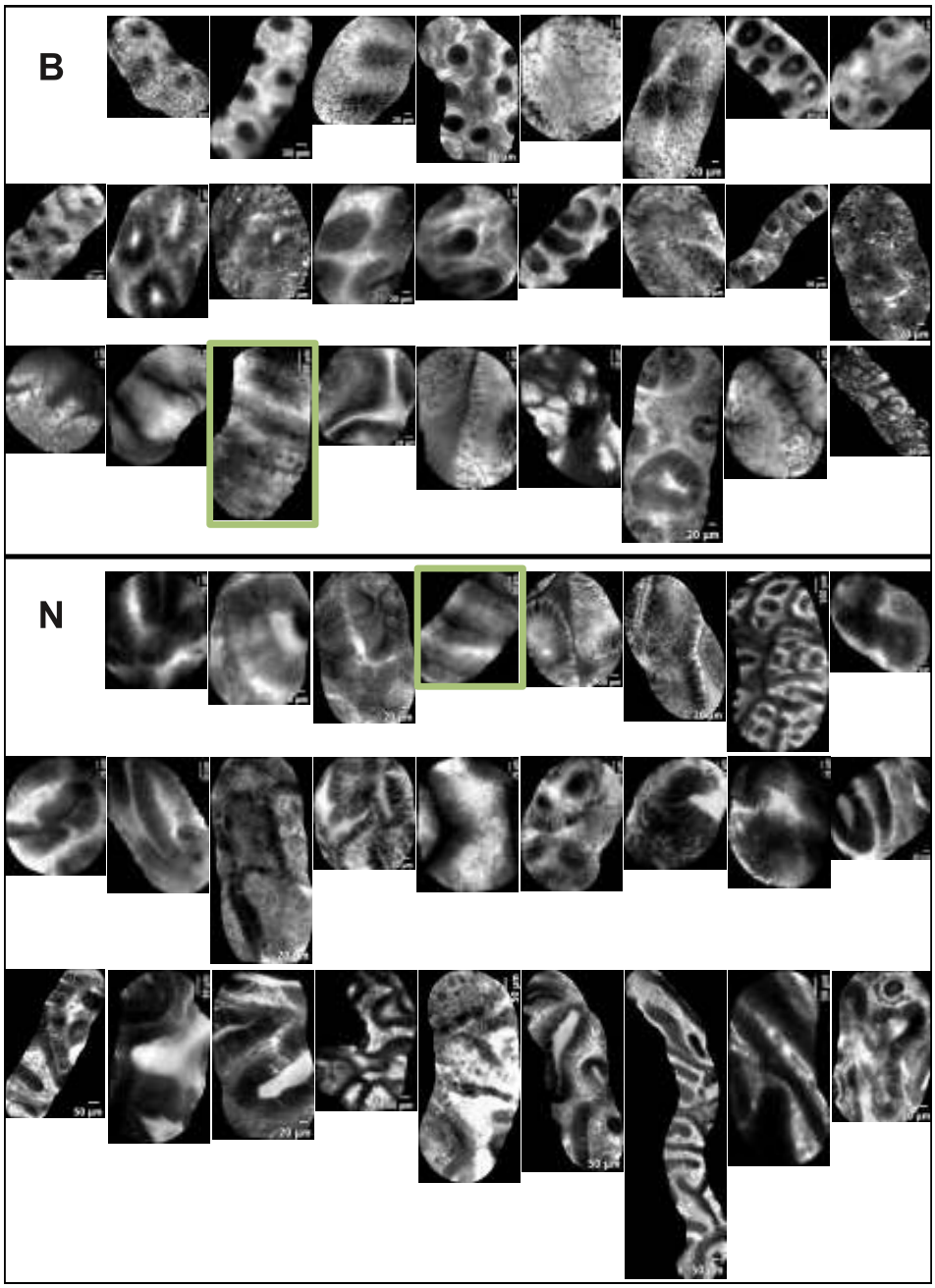


Figure 2.2: pCLE mosaic samples from our database of colonic polyps. These mosaics were built from image sequences using a video-mosaicing technique with non-rigid registration [Vercauteren 06]. Scale bars provide a cue on the field of view size. On top (resp. bottom) are the mosaics of the polyps diagnosed as non-neoplastic (resp. neoplastic) indicated by **B** (resp. **N**). The closer to the boundary the mosaics are, the less obvious is their diagnosis according to their visual appearance. The two framed mosaics might look similar although they belong to different pathological classes.

Currently, pCLE is relatively new to many physicians, who are still in the process of defining a taxonomy of the pathologies seen in the image sequences. To support the endoscopist in establishing a diagnosis, we aim to extract, from a training database, endomicroscopic videos that have a similar appearance to a video of interest but have been previously annotated by expert physicians with a textual diagnosis confirmed by histology. Our main objective is Content-Based Image Retrieval (CBIR) applied to pCLE videos. However, it is difficult to have a ground truth for CBIR, because of the subjective appreciation of visual similarities. An objective indirect method to evaluate retrieval performance is classification. In our approach, we make a clear distinction between retrieval, which is the target of this study, and classification, which is the indirect means that we choose to evaluate the retrieval performance. For didactic purposes, we explore the image retrieval approach as a first step and we then move progressively to video retrieval which is our final goal.

In the clinical field, the important need for medical image retrieval has been clearly expressed in the scientific literature, for example by Long et al. [Long 09], Müller et al. [Müller 04], and Smeulders et al. [Smeulders 00]. Particularly, the medical image retrieval task of "ImageCLEF", presented in [Müller 08], proposes a publicly-available benchmark for the evaluation of several multimodal retrieval systems. However the application of retrieval for endomicroscopy has not yet been investigated. Histological images are the closest in appearance to pCLE images. In histology analysis, many efforts have been made to automate pathological differentiation: by Gurcan et al. in [Gurcan 09], by Kong et al. in [Kong 09], or by Doyle et al. in [Doyle 06]. Recently, the "PR in HIMA" Contest, launched in 2010, addresses the issue of pattern recognition in digital histology images. Nevertheless, many standard computer-aided diagnosis features that are commonly employed in histology image analysis cannot be used in our retrieval application because they are simply not visible. For example, the nuclear-cytoplasmic ratio cannot be computed because nuclei and membranes are hardly visible in pCLE images.

Observing that epithelial tissues are characterized by the regularity of the cellular and vascular architectures, our objective is to retrieve discriminative texture information coupled with shape information by applying local operators on pCLE images. To serve that purpose, we revisit in Section 2.3 the Bag-of-Visual-Words (BoW) method, proposed by Sivic and Zisserman [Sivic 06], which has been successfully used in many applications of computer vision: from the categorization of textures and objects, as presented by Zhang et al. [Zhang 07], to the recognition of human actions in movies, as presented by Laptev et al. [Laptev 08]. To apprehend the large intra-class variability of our pCLE database, we refer the reader to Fig. 2.2, where single images of colonic polyps belong to either neoplastic epithelium, i.e. the *pathological* class, or non-neoplastic epithelium, i.e. the *benign* class. We can also observe small inter-class differences: Two pCLE images may have a quite similar appearance but with an opposite diagnosis. We looked at describing discriminative information in pCLE images, by taking into account the physics of the acquisition process explained in Section 2.2.1, as well as the type of invariance

necessary for their retrieval. By adjusting the image description to these invariants in Section 2.3, we were able to considerably improve the retrieval and provide more relevant similar images. Our other main adjustments consist of choosing a dense detector that captures the densely distributed information in the image field, similarly to what was proposed by Leung and Malik [Leung 01] with texture patches, and performing a local multi-scale image description that extracts microscopic as well as mesoscopic features.

Because the field of view (FoV) of single images may not be large enough to perform a robust diagnosis, expert physicians focus in practice on several images for the interpretation. To solve the FoV problem but still be able to work on images rather than videos, we consider, as objects of interest for the retrieval, larger mosaic images that are built from the image sequences using the video-mosaicing technique of Vercauteren et al. [Vercauteren 06]. The high degree of variability in appearance also holds for the resulting mosaic images, as shown in Fig. 2.1. To improve the state of the art in CBIR, we define an efficient similarity metric based on the visual words, taking into account their discriminative power with respect to the different pathological classes. One intrinsic limitation of the standard BoW representation of an image is that spatial relationships between local features are lost. However, as the spatial organization of cells is highly discriminative in pCLE images, we aim at measuring a statistical representation of this geometry. By exploiting the co-occurrence matrix of visual words, we extract a geometrical measure that is applied after the retrieval to remove possible outliers.

Building mosaic images using non-rigid registration tools requires a substantial amount of time, which is undesirable for supporting diagnosis in near real-time. In Section 2.5, to reach the objective of interactive CBIR, we take advantage of the coarse registration results provided by the real-time mosaicing proposed by Vercauteren et al. [Vercauteren 08]. We include, in the retrieval process, the possible spatial overlap between the images from the same video sequence. A histogram summation technique also reduces retrieval runtime.

The binary classification results show that our retrieval method achieves substantially better accuracy than several state-of-the art methods, and that using video data provides a statistically significant improvement when compared to using single images independently. A finer retrieval evaluation based on multi-class classification is proposed in Section 2.6, with encouraging results.

2.2 Context of the Study

2.2.1 Probe-based Confocal Laser Endomicroscopy

During an ongoing endoscopy procedure, pCLE consists of imaging the tissue at microscopic level, by inserting, through the standard endoscope, a miniprobe made of tens of thousands of optical fibers. A proximal part laser scanning unit uses two mirrors to emit, along each fiber, an excitation light that is locally absorbed by fluorophores in the tissue; the light which is then emitted by the fluorophores

at a longer wavelength is transferred back along the same fiber to a mono-pixel photodetector, as illustrated in Fig. 2.3. As a result, endomicroscopic images are acquired at a rate of 9 to 18 frames per second, composing video sequences. From the irregularly-sampled images that are acquired, an interpolation technique presented by Le Goualher et al. [Le Goualher 04] produces single images of diameter 500 pixels, which corresponds to a FoV of 240 μm, as illustrated in Fig. 2.6. All the pCLE video sequences that are used for this study have been acquired by the Cellvizio system of Mauna Kea Technologies.

Considering a video database of colonic polyps, our study focuses on supporting the early diagnosis of colorectal cancers, more precisely for the differentiation of neoplastic and non-neoplastic polyps.

2.2.2 Endomicroscopic Database

At the Mayo Clinic in Jacksonville, Florida, USA, 68 patients underwent a surveillance colonoscopy with pCLE for fluorescein-aided imaging of suspicious colonic polyps before their removal. For each patient, pCLE was performed of each detected polyp with one video corresponding to each particular polyp. All polyps were removed and evaluated by a pathologist to establish the "gold standard" diagnosis. In each of the acquired videos, stable sub-sequences were identified by clinical experts to establish a diagnosis. They differentiate pathological patterns from benign ones, according to the presence or not of neoplastic tissue which contains some irregularities in the cellular and vascular architectures. The resulting *Colonic Polyp* database is composed of 121 videos (36 benign, 85 neoplastic) split into 499 video sub-sequences (231 benign, 268 neoplastic), leading to 4449 endomicroscopic images (2292 benign, 2157 neoplastic). For all the training videos, the

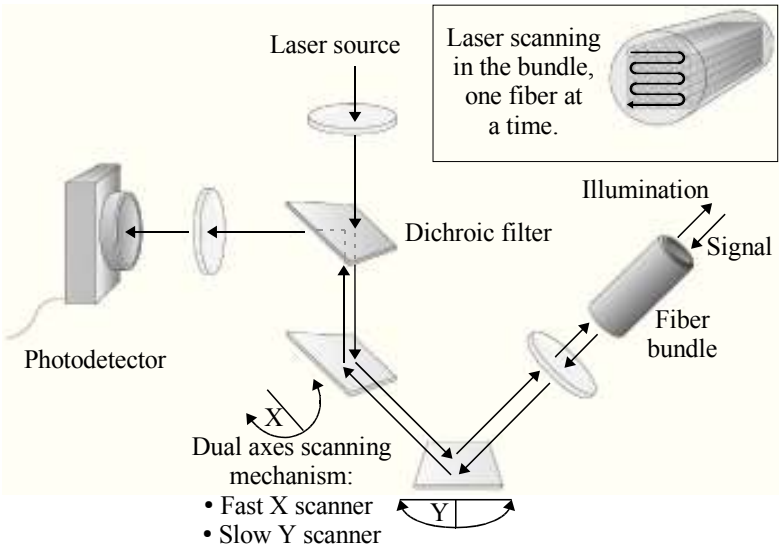


Figure 2.3: Physics of acquisition for pCLE imaging.

pCLE diagnosis, either benign or neoplastic, is the same as the "gold standard" established by a pathologist after the histological review of biopsies acquired on the imaging spots.

More details about the acquisition protocol of the pCLE database can be found in the studies of Buchner et al. [Buchner 08], [Buchner 09b], which included a video database of colonic polyps comparable to ours, and demonstrated the effectiveness of pCLE classification of polyps by experts endoscopists.

2.2.3 State-of-the-Art Methods in CBIR

In the field of computer vision, Smeulders et al. [Smeulders 00] presented a large review of the state of the art in CBIR. At the macroscopic level, Häfner et al. [Häfner 09] worked on endoscopic images of colonic polyps and obtained rather good classification results by considering 6 pathological classes. At the microscopic level, Désir et al. [Désir 10] investigated the classification of pCLE images of the distal lung. However, the goal of these two studies is classification for computer-aided diagnosis, whereas our main objective is retrieval. Petrou et al. [Petrou 06] proposed a solution for the description of irregularly-sampled images, which could be defined by the optical fiber positions in our case. Nevertheless, we chose for the time being not to work on irregularly-sampled images, but rather on the interpolated images, for two reasons: first, we plan to retrieve pCLE mosaics which are interpolated images, and second, most of the available retrieval tools from computer vision are based on regular grids. The following paragraphs present several state-of-the-art methods that can be easily applied to endomicroscopic images and that will be used as baselines in this study to assess the performance of our proposed solutions.

In addition to the BoW method presented by Zhang et al. [Zhang 07] which is referred to as the HH-SIFT method combining sparse feature extraction with the BoW model, we will take as references the two following methods for CBIR method comparison: first, the standard approach of Haralick features [Haralick 79] based on global statistical features and experimented by Srivastava et al. [Srivastava 08] for the identification of ovarian cancer in confocal microendoscope images, and second, the texture retrieval Textons method of Leung and Malik [Leung 01] based on dense local features.

The Haralick method computes global statistics from the co-occurrence matrix of the image intensities, such as contrast, correlation or variance, in order to represent an image by a vector of statistical features; this method is worth being compared with, because of its global scope. The Textons method defines for each image pixel p a "texton", as the response of a patch centered on p to a texture filter which is composed of orientation and spatial-frequency selective linear filters. While only texture information is extracted by this method, the fact that its extraction procedure is dense makes it interesting for method comparison, as shown in Section 2.3.

2.2.4 Framework for Retrieval Evaluation

Assessing the quality of content-based data retrieval is a difficult problem. In this paper, we focus on a simple but indirect means to quantify the relevance of retrieval: we perform classification. We chose one of the most straightforward classification method, the k-nearest neighbors (k-NN) method, even though any other method could be easily plugged in our framework. We first consider two pathological classes, benign ($C = -1$) and neoplastic ($C = +1$), then we propose a multi-class evaluation of the retrieval in Section 2.6. As an objective indicator of the retrieval relevance, we take the classification accuracy (number of correctly classified samples / total number of samples).

In order to determine if the improvement from one retrieval method to another is statistically significant, we will perform the McNemar's test [Sheskin 11] based on the classification results obtained by the two methods at a fixed number of nearest neighbors. We refer the reader to the Appendix A for a detailed description of the McNemar's test.

Given the small size of our database, we need to learn from as much data as possible. We thus use the same database both for training and testing but take great care into not biasing the results. If we only perform a leave-one-out cross-validation, the independence assumption is not respected because several videos are acquired on the same patient. Since this may cause bias, we chose to perform a leave-one-patient-out (LOPO) cross-validation, as introduced by Dundar et al. [Dundar 04]: All videos from a given patient are excluded from the training set before being tested as queries of our retrieval and classification methods. Even though we tried to ensure unbiased processes for learning, retrieval and classification, it might be argued that some bias is remaining because splitting and selection of video subsequences were done by one single expert. For our study we can consider this bias as negligible.

It is worth mentioning that, in the framework of medical information retrieval, some scenarios require predefined sensitivity or specificity goals, depending on the application. Some applications, such as brain surgery, may require a predefined high specificity. For our application, physicians prefer to have a false positive caused by the misdiagnosis of a benign polyp, which could lead for example to unnecessary but well supported polypectomy, than to have a false negative caused by the misdiagnosis of a neoplastic polyp, which may have serious consequences for the patient. Thus, our goal is to reach the predefined high sensitivity, while keeping the highest possible specificity. For this reason, we introduce a weighting parameter $\theta \in [-1, 1]$ to trade-off the cost of false positives and false negatives. Given the pathological classes $C^{j\in\{1,..,k\}} \in \{-1, +1\}$ of the k nearest neighbors and the similarity distances $d^{j\in\{1,..,k\}}$ from them to the query, the query is classified as neoplastic if and only if:

$$\frac{\sum_{j=1}^{k} \frac{C^j}{d^j}}{\sum_{j=1}^{k} \frac{1}{d^j}} \; > \; \theta \tag{2.1}$$

The default value of the additive threshold θ is $\theta = 0$, which corresponds to the situation where the pathological votes of all the k neighbors have the same weight. Negative values of θ correspond to putting more weight to neoplastic votes. The closer θ is set to -1 (resp. $+1$), the more weight we give on the neoplastic votes (resp. the benign votes) and the larger the sensitivity (resp. the specificity) is. ROC curves can thus be generated by computing the couple (specificity, sensitivity) at each value of $\theta \in [-1, 1]$, which provides another way to evaluate the classification performance of any of the retrieval methods.

One may argue that our methodology uses an ad-hoc number of visual words and is thus dependent on the clustering results. This is the reason why, in Section 2.5.2, we will compare the classification performances of our retrieval method with those of a simple yet efficient image classification method, the Naive-Bayes Nearest-Neighbor (NBNN) classifier of Boiman et al. [Boiman 08], that uses no clustering but was proven to outperform BoW-based classifiers. For each local region of the query the NBNN classifier computes, in the description space, its distances respectively to the closest region of the benign and neoplastic training data sets. If the sum of the benign distances D_B is smaller than the sum of the neoplastic distances D_N, the query is classified as benign, otherwise as neoplastic [Boiman 08]. The construction of ROC curves for the NBNN classification method requires the use of a multiplicative threshold $\theta_{\text{NBNN}} \in [0, +\infty[$ according to which the query is classified as neoplastic if and only if:

$$D_N \; < \; \theta_{\text{NBNN}} \, D_B \tag{2.2}$$

The default value of the multiplicative threshold θ_{NBNN} is $\theta_{\text{NBNN}} = 1$, which corresponds to the situation where the pathological votes of all the k neighbors have the same weight. Values of θ_{NBNN} greater than 1 correspond to putting more weight to neoplastic votes. The larger (resp. smaller) θ_{NBNN} is set, the more weight we give on the neoplastic votes (resp. the benign votes) and the larger the sensitivity (resp. the specificity) is.

Another characteristic of our application is that pCLE videos diagnosed as neoplastic may contain some benign patterns whereas benign epithelium never contains neoplastic patterns. Therefore, it seems logical to put more weight on the neoplastic votes, being more discriminative than benign votes. The weighting parameters θ and θ_{NBNN} may also be useful to compensate for our unbalanced dataset, which contains more benign images than pathological ones.

2.3 Adjusting Bag-of-Visual-Words for Endomicroscopic Images

2.3.1 Standard Bag-of-Visual-Words Method

As one of the most popular recent methods for image retrieval, the standard BoW method consists of detecting salient image regions from which continuous features

are extracted and discretized. All features are clustered into a finite number of bins called "visual words", whose number of occurrences in an image constitute the image signature. As illustrated in Fig. 2.4, the BoW retrieval process can thus be decomposed into four main steps: salient region detection, region description, description vectors clustering, and similarity measurement based on the signatures. After the description step, the image is typically represented in a high-dimensional space by a set of description vectors. To reduce the dimension of the description space, a standard K-Means clustering step builds K clusters, from the union of the description vector sets gathered across all the images of the training database. K visual words are then defined, each one being the mean of a cluster in the description space. Each description vector counts for one visual word, and one image is represented by a signature of size K which is its histogram of visual words, normalized by the total number of local regions. Given the image signatures, the similarity distance between two images can be defined as an appropriate distance between their signatures.

The advantage of the simple metric provided by the χ^2 pseudo-distance is that it is only based on the comparison between the values within the same histogram bin. If $H_I = (w_1^I, ..., w_K^I)$ and $H_J = (w_1^J, ..., w_K^J)$ are the histograms of the two images I and J, where $w^I(i)$ is the frequency of the i^{th} visual word in the image I,

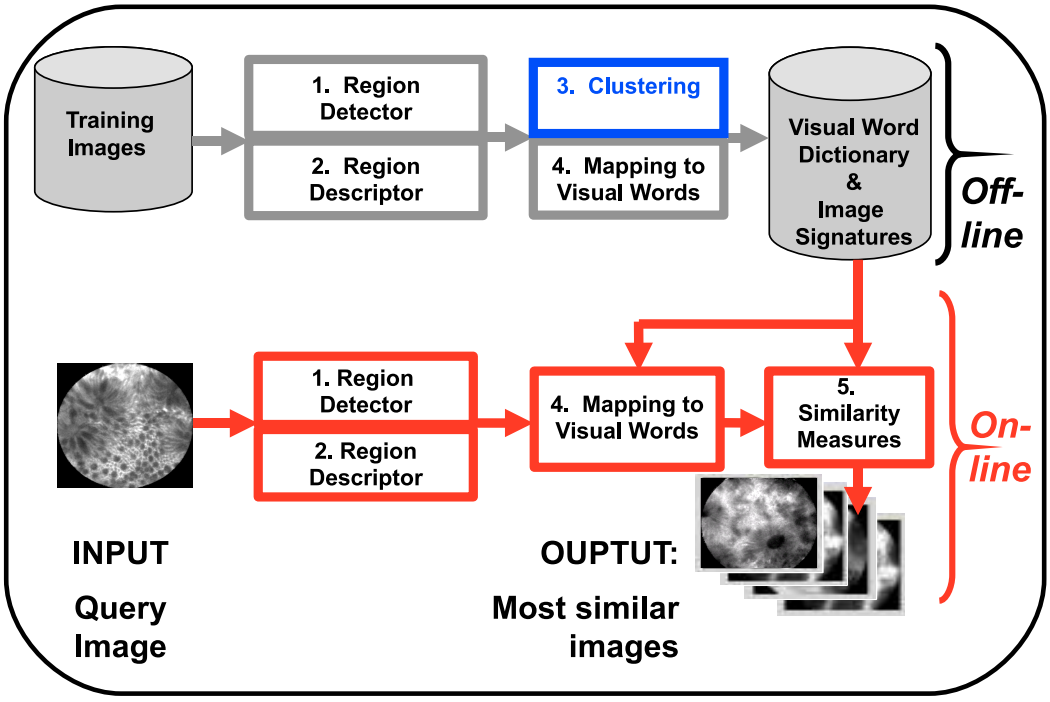


Figure 2.4: Overview of the retrieval pipeline.

then the similarity distance between I and J is defined as:

$$\chi^2(H_I, H_J) = \frac{1}{2} \sum_{i=1, w_i^I w_i^J > 0}^{K} \frac{(w_i^I - w_i^J)^2}{w_i^I + w_i^J} \tag{2.3}$$

In these conditions, as explained by Sivic and Zisserman [Sivic 06], similarity measurement is quite efficient and can be approximated by the term frequency - inverse document frequency (TF-IDF) technique for a fast retrieval runtime. Nister and Stewenius [Nister 06] showed that, combined with a hierarchical clustering, the inverted file indexing enables large-scale data retrieval. With the same purpose of k-NN approximation, Muja and Lowe [Muja 09] developed the Fast Library for Approximate Nearest Neighbors (FLANN) available in OpenCV. Among the more sophisticated metrics, the Earth Mover's Distance (EMD) proposed by Rubner et al. [Rubner 00] may be more relevant than the χ^2 metric because it accounts for the full vector representation of the visual words in the SIFT space. But standard EMD is less computationally efficient than χ^2 because it needs to compute distances in high-dimensional space in order to calculate the transportation costs from one visual word to another. Nevertheless, it would be interesting to test the fast implementation of EMD that has been recently presented by Pele and Werman [Pele 09]. For the classification step that quantifies the similarity results, the votes of the k-nearest neighbors can be weighted by the inverse of their χ^2 pseudo-distance to the tested image signature, so that the closest images are the most discriminant.

Recognized as a powerful feature extraction method in computer vision, the HH-SIFT method uses the Harris-Hessian (H-H) detector coupled with the Scale Invariant Feature Transform (SIFT) descriptor proposed by Lowe [Lowe 04]. When applied to the non medical UIUCTex database of textures, which is admittedly a rather easy database, the HH-SIFT method of Zhang et al. [Zhang 07] achieves excellent retrieval results and yields a classification accuracy close to 98% for 25 texture classes. However, when we applied this method, as well as other state-of-the-art methods, on our pCLE database, we obtained rather poor retrieval results and we observed the presence of many outliers in the retrieval. This was confirmed by the associated low classification results presented in Fig. 2.7: when considering only 2 classes, the accuracy is below 67%, which is not acceptable for clinical use. We will show that even though the standard BoW method is not adapted for the retrieval of endomicroscopic images, the adjustments that we propose can turn it into a powerful tool for our needs. For instance by taking into account the pCLE imaging system, we can leverage the constraints that characterize our retrieval application. Our first contributions are presented in Sections 2.3 and 2.4. We explored them in a preliminary study [André 09a].

2.3.2 Moving to Dense Detection of Local Regions

It is worth noticing that the endoscopists examine, in the colonic epithelium, goblet cells and crypts which are round-shaped or tubular-shaped, as illustrated in Fig. 2.6.

For this reason, we first looked at extracting blob features in the images by applying sparse detectors. Sparse detectors extract salient regions in the image, i.e. regions containing some local discriminative information. In particular, the H-H operator detects corners and blobs around key-points with high responses of intensity derivatives for at least two distinct gradient directions. Other sparse detectors like the Intensity-Based Regions (IBR) of Tuytelaars and Van Gool [Tuytelaars 00] and the Maximally Stable Extremal Regions (MSER) of Matas et al. [Matas 04] are also specialized for the extraction of blob features.

However, while testing on pCLE videos the numerous sparse detectors listed in [Mikolajczyk 05, Tuytelaars 08], we observed that a large number of salient regions do not persist between two highly correlated successive images taken from the same video, as shown in Fig. 2.5. In fact, these detectors have been designed for computer vision applications and seem to be inadequate for our medical application because of their sparse nature: they fail to capture all the discriminative information which is densely distributed in pCLE images. This may explain the poor retrieval results on pCLE images of the HH-SIFT method, which uses the sparse H-H detector.

To capture all the interesting information, we decided to apply a dense detector made of overlapping disks of constant radius. These disk regions are localized on a regular grid, such that each disk covers a possible image pattern at a microscopic level, as illustrated in Fig. 2.6. With the regular dense operator, we will show already promising results in the following section. The benefits of a dense operator for image retrieval have also been demonstrated with the pixel-wise approach of "TextonBoost" by Shotton et al. [Shotton 06], who were mainly interested in object categorization and segmentation problems.

Using the BoW method with dense detection enables the dense visualization of visual words on the entire image field. In each described image, we decided

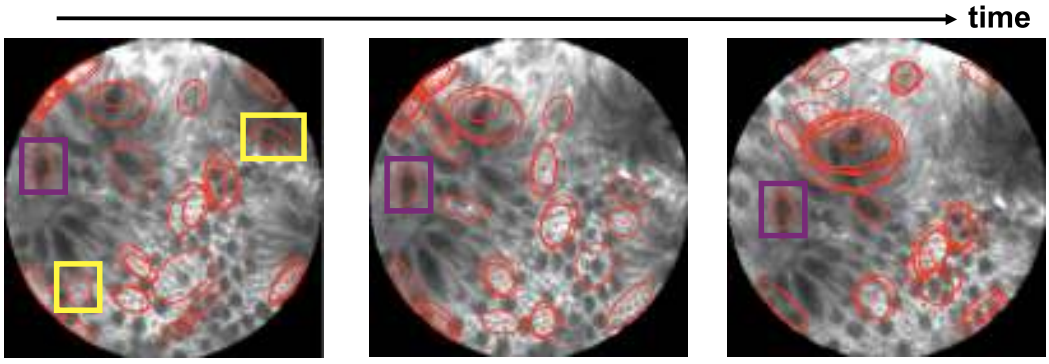


Figure 2.5: Salient regions (ellipses) extracted by the sparse MSER detector on three successive frames of a benign video sequences. Some regions, like the one framed in dark, are correctly followed by the detector, but many others, like those framed in bright, are lost. This shows the inconsistency of the sparse detector for the description of pCLE images.

to map the visual words to different colors and to superimpose on the image the local disk regions filled with the color of their visual word index. In the description space, the relative distances between the visual words is missing in their arbitrary numbering after the clustering process. As we wanted the colors to convey a feeling on these distances, we decided to project the high-dimensional clusters representing the visual words onto the three-dimensional RGB space, using a simple Principal Component Analysis (PCA). Then, each of the K visual words is mapped to a specific color. As a result, the superimposed colors are highlighting the geometrical structures in the images, as illustrated in Fig. 2.17.

For qualitative interpretation, we wanted to be able to visually compare the spatial distributions of the visual words in two image queries that may come from different patients. So, for the *display* of the colored visual words *only*, we did not apply the LOPO procedure according to which, for each patient, a different clustering process must be done that excludes the patient from the training dataset and generates different visual words. Instead, we generated the $K = 100$ visual words only once for the visualization, by performing a single clustering process on the total number of SIFT vectors that describe the images associated with all the patients of the database.

2.3.3 Multi-Scale Description of Local Regions

Let us now look at what kind of invariants are necessary for the description of pCLE images. The distance of the probe's optical center to the tissue does not change while imaging, so the only possible motions of the pCLE probe along the tissue surface are translations and in-plane rotations. For this reason, we aim at describing pCLE images in an invariant manner with respect to translation and in-plane rotation. Besides, as the rate of fluorescein injected before imaging procedure is decreasing through time, we want this description to be also reasonably invariant to intensity changes. For this purpose, the standard SIFT description appeared to be the most appropriate since it extracts a local image description which, when coupled with an invariant detector, is invariant to affine transformations of the intensity and some viewpoint changes, e.g., translations, rotations and scaling. Indeed, the SIFT descriptor computes, for each salient region, a 128-bin description vector which is its gradient histogram at the optimal scale provided by the detector, the gradient orientations being normalized with respect to the principal orientation of the salient region. We refer the reader to the study of Zhang et al. [Zhang 07] for a survey of the SIFT descriptor or other powerful ones. In particular, the Speeded Up Robust Features (SURF) descriptor of Bay et al. [Bay 06] is more efficient than SIFT in terms of runtime, but was not considered in this study.

There is no scale change in the pCLE imaging system because the distance from the probe to the tissue is fixed: a given clinical pattern should have the same scale in all the images in which it is present. In colonic polyps, however, mesoscopic crypts and microscopic goblet cells both have a rounded shape, but are different objects characterized by their different sizes. This is the reason why we need a scale

dependent description, instead of the standard scale invariant description. In order to capture information at different scales, we define local disk regions at various scales using fixed values, for example by choosing a microscopic scale for individual cell patterns and a mesoscopic scale for larger groups of cells. This leads us to represent an image by several sets of description vectors that are scale-dependent, resulting in several signatures for the image that are then concatenated into one larger signature.

For our experiments on the dense description, we considered disk regions of radius 60 pixels to cover groups of cells. We then chose 20 pixels of grid spacing to get a reasonable overlap between adjacent regions and thus be nearly invariant with respect to translation. Besides, among the values from 10 to 30000 that we found in the literature for the number K of visual words provided by the K-Means clustering, the value $K = 100$ yielded satisfying classification results on our relatively small database. The classification results that quantify the retrieval of single images are presented in Fig. 2.7 where we observe that, compared to the standard HH-SIFT method, the dense detector brings a gain of accuracy of 17.1 percentage points at $k = 10$ neighbors, with a resulting accuracy of 81.7% (78.0% sensitivity, 85.1% specificity). The McNemar's tests show that, with statistical significance, our dense method is better than the other methods (p-value $< 10^{-6}$ for $k \in [1, 10]$), Texton is better than Haralick (p-value < 0.0040 for $k \in [1, 10]$), and Haralick is better than HH-SIFT (p-value $< 10^{-6}$ for $k \in [1, 10]$).

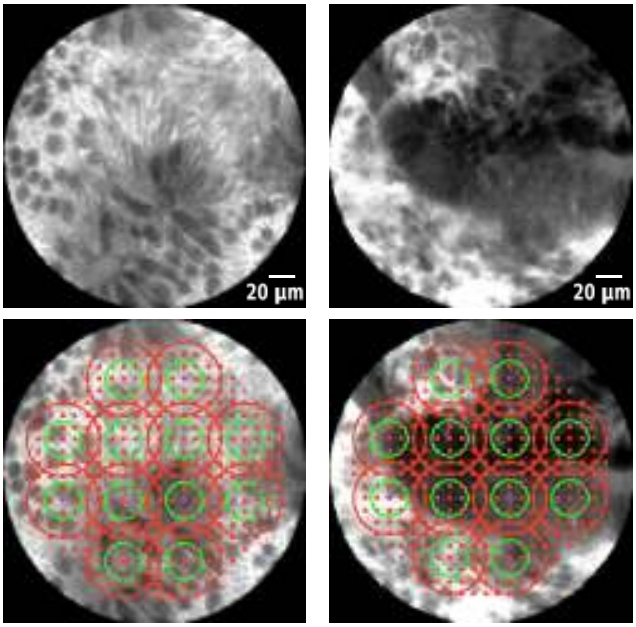


Figure 2.6: Small and large disk regions on a dense regular grid, applied on a benign image (left), and on a neoplastic image (right). Small disks of radius 30 pixels cover microscopic information like individual cells, whereas large disks of radius 60 pixels cover mesoscopic information like groups of cells. The images have a diameter of approximately 500 pixels that corresponds to a FoV of 240 μm.

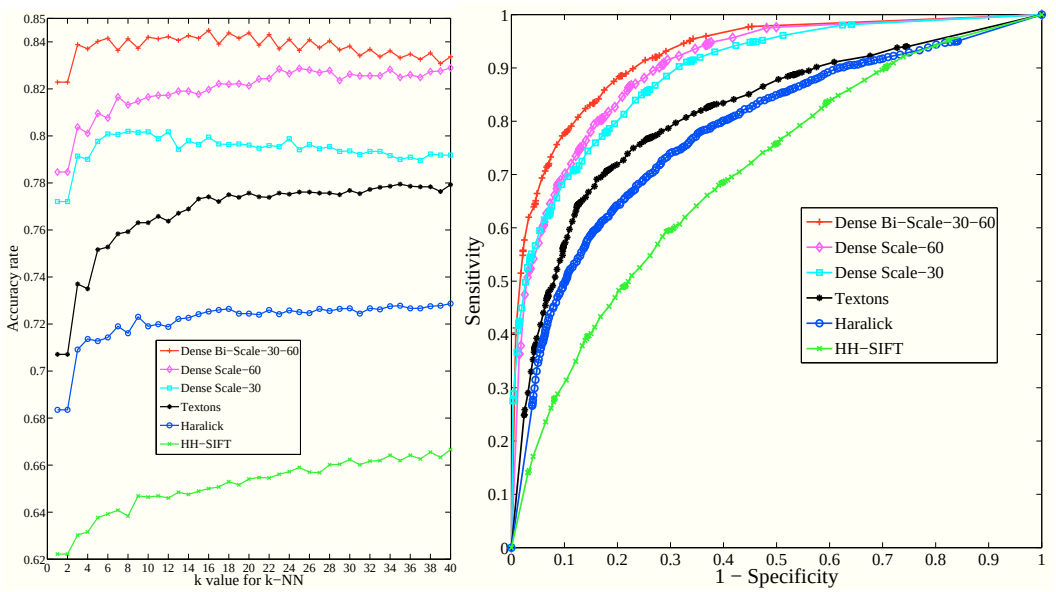


Figure 2.7: Left: LOPO classification of single pCLE images by the methods, with the default value $\theta = 0$. Right: Corresponding ROC curves at $k = 10$ neighbors with $\theta \in [-1, 1]$. θ trades off the cost of false positives and false negatives.

For our experiments on the bi-scale description, a large disk radius of $\rho_1 = 60$ pixels is suitable to cover groups of cells, while a smaller disk of radius $\rho_2 = 30$ pixels allows to cover at least one cell in the images, as shown in Fig. 2.6. For the classification of single images, we observe in Fig. 2.7 that, when compared to the one-scale description of the Dense-Scale-60 (D-S-60) method, the bi-scale description of the Dense-Bi-Scale-30-60 (D-BS-30-60) method brings an additional gain of accuracy of 2.5 percentage points at $k = 10$ neighbors, with a resulting accuracy of 84.2% (80.8% sensitivity, 87.4% specificity). Besides, McNemar's tests show that this classification improvement is statistically significant (p-value $< 10^{-6}$ for $k \in [1, 10]$), thanks to the complementarity of our two scale-dependent descriptors.

2.4 Contributions to the State of the Art

2.4.1 Solving the Field-of-View Issues using Mosaic Images

In the retrieved single images, we often observed single images with a similar appearance to the query but attached to the opposite diagnosis. One important reason is that, on a single pCLE image, some discriminative patterns, e.g. an elongated crypt, may only be partially visible and so unable to characterize the pathology. To address this FoV issue, we aimed at performing the retrieval beyond single images. In our pCLE video database, the dynamic motion within the tissue can be neglected when compared to the global motion of the probe sliding along the tissue surface. In stable video sequences, the miniprobe is in constant contact with the tissue,

so the distance of the probe's optical center to the tissue is fixed. As successive images from the same video are mostly related by viewpoint changes, we can use the video-mosaicing technique of Vercauteren et al. [Vercauteren 06], to project the temporal dimension of a video sequence onto one mosaic image with a larger FoV and of higher resolution. Even if time information is lost after the mosaicing, Becker et al. [Becker 07] showed that the mosaic image produced by this video-mosaicing technique has a clinical interest in endomicroscopy. Several applications of video-mosaicing as a support for pCLE interpretation have been presented, for example by De Palma et al. [De Palma 10] on the colon and by Thiberville et al. [Thiberville 07] on the lung.

Thus, instead of single images, we considered mosaic images as objects of interest for the retrieval. All videos of the database were first split into stable video sub-sequences identified by expert physicians. These stable subsequences remain after the removal of unreliable parts of the videos that correspond either to fast motions of the probe leading to motion artifacts, or to the moments when the probe has lost contact with the tissue. Then we built mosaics on these video sub-sequences and we applied the dense BoW method directly on the produced mosaic images. As the discriminative information that we extracted in the single images is kept in the mosaic images, we chose the same values of parameters for the radii of 30 and 60 pixels of the disk regions and for the number $K = 100$ of visual words. However, as larger discriminative patterns may be present in mosaic images, we thought that larger scale features should capture them. For this purpose, we evaluated, without cross-validation as a first step, mosaic retrieval using successively the D-S-80 method (dense regions of radius 80 pixels), the D-S-100 method (dense regions of radius 100 pixels), and the D-BS-60-80 method that concatenates the mosaic signatures of D-S-60 and D-S-80. The classification results without cross-validation showed that D-S-80 and D-BS-60-80 are comparable to D-S-60, and that D-S-100 performs worse than D-S-60. For this reason, we decided to evaluate only D-S-30, D-S-60 and D-BS-30-60 with LOPO cross-validation. We think that a reason why larger scale features fail to capture larger discriminative patterns in mosaic images may be the trade-off between smoothing and region size in the SIFT description. Besides, the larger the size of the regions is, the more discriminative the shape of the regions is in the image description, and our circular-shaped regions may not be adequate anymore. Indeed, at scales larger than 60 pixels of radius, ellipsoidal regions should better capture elongated patterns such as abnormal crypts.

The accuracy results for the classification of mosaic images are presented in Fig. 2.8. They show that the compared retrieval methods follow the same order of performance as the one we observed on single images. Besides, our dense retrieval methods achieve more satisfying classification results for the retrieval of mosaic images than for the retrieval of single images. With statistical significance, D-S-60 is better than Texton (p-value $< 10^{-6}$ for $k \in [1, 10]$), and Texton is better than Haralick (p-value < 0.0057 for $k \in [1, 2]$). For $k \in [1, 10]$ the performances of Haralick and HH-SIFT are comparable; for more neighbors, Haralick outperforms HH-SIFT with statistical significance (p-value < 0.032 for $k \in [15, 20]$). However, for

the comparison between D-S-60 and D-BS-30-60 (p-value ≥ 0.11 for $k \in [1, 10]$), the performance differences are not statistically significant. A possible reason is that, on our database, all the discriminative information may have already been captured at the scale 60. We hope that, with a larger pCLE database, the bi-scale description will significantly improve retrieval performance. The best result for the classification of mosaic images is reached by the dense bi-scale description method denoted by D-BS-30-60, at $k = 6$ neighbors, with an accuracy of 88.2% (sensitivity 91.0%, specificity 84.9%). These results are close to the clinical expectations. Nevertheless, we will show that we can still improve them for our clinical application.

2.4.2 Similarity Metric based on Visual Words

The similarity metric defined by the χ^2 pseudo-distance is efficient but highly sensitive to the frequency of each visual word in an individual image with respect to its frequency in the whole set of images. More importantly, the ability of the retrieved images to represent the pathological class of the query is thus sensitive to the discriminative power of the visual words with respect to the pathological classes.

To address this problem, we propose to weight, according to their discriminative power, the contributions of the visual word frequencies to the metric. For each class $C \in \{-1, +1\}$ of images, we considered the distribution $p(i|C)$ of the frequencies of the i^{th} visual word in the images belonging to the class C. We define the discriminative power $g(i)$ of the i^{th} visual word using the Fisher criterion between the two

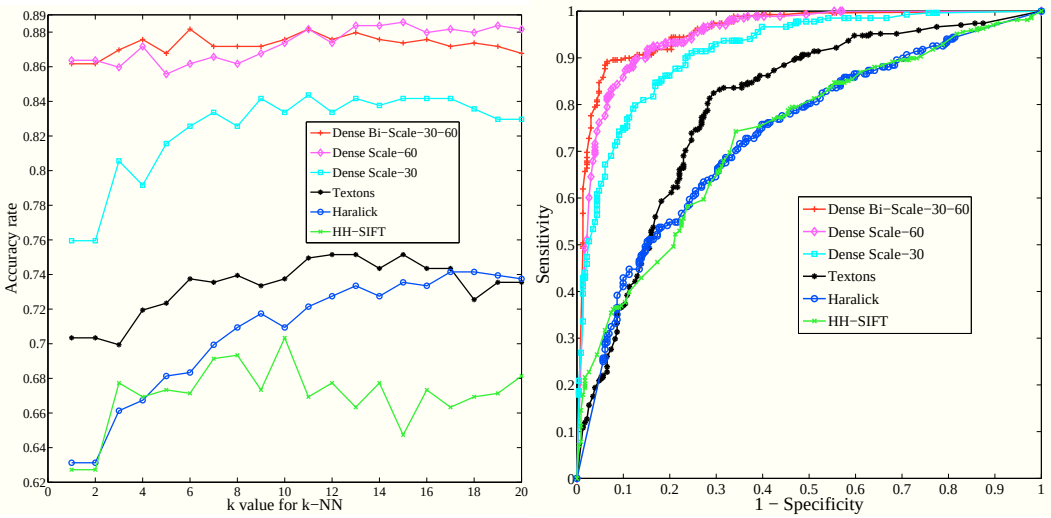


Figure 2.8: Left: LOPO classification of pCLE mosaic images by the methods, with the default value $\theta = 0$. Right: Corresponding ROC curves at $k = 5$ neighbors with $\theta \in [-1, 1]$. θ trades off the cost of false positives and false negatives. The mosaic images have been built with non-rigid registration.

distributions $p(i|C=-1)$ and $p(i|C=+1)$:

$$g(i) = \frac{(\mu_{-1}(i) - \mu_{+1}(i))^2}{\sigma_{-1}(i)^2 + \sigma_{+1}(i)^2} \tag{2.4}$$

where $\mu_C(i)$ and $\sigma_C(i)^2$ are respectively the mean and the variance of the distribution of the frequencies of the i^{th} visual word in the images belonging to class C, with $C \in \{-1, +1\}$. Our approach, that combines L^1-normalization applied to the visual word histograms, and Fisher weighting applied to the visual words, could be composed with other similarity metrics than χ^2, some of which are presented in [Sivic 09]. Besides, it is close to other approaches exploiting discriminative context information, such as the TF-IDF technique, or the Fisher kernels method which is used by Perronnin and Dance [Perronnin 07] as an extension of the BoW method for image categorization. Discriminative vocabulary learning was also investigated by Winn et al. [Winn 05], who proposed pair-wise merging of visual words to derive an original statistical measure of the discriminative power.

An alternative to visual word weighting is the selection of the most discriminative visual words, i.e. those minimizing the intra-class distances while maximizing the inter-class distances. This corresponds to a binary weighting, which decreases the size of image signatures by reducing the number of visual words, so the image retrieval and classification processes run faster. For our experiments, the K' most discriminative visual words are selected from the $K = 100$ original ones by applying on their discriminative power a threshold λ. Changing the value of λ may have an influence on the classification accuracy based on these signatures. After testing the whole training set without cross-validation we chose $\lambda = 0.7$, so that 20% to 25% of the visual words are selected, which ensures both significantly shorter signatures and better classification accuracy. This threshold λ is applied inside each cross-validation sub-set for which it selects a certain number of discriminative visual words. The mean value of K' over all cross-validation sub-sets is 23.2.

The classification of mosaic images presented in Fig. 2.9 shows that, coupled with the dense detector and the biscale description, the visual word binary *selection* brings an additional gain of accuracy of 2.0 percentage points at $k = 5$ neighbors, with a resulting accuracy of 88.8% (91.0% sensitivity, 86.2% specificity). Although we established that this classification improvement is not statistically significant (p-value ≥ 0.15 for $k \in [1, 10]$), the binary selection reduces retrieval runtime while reaching comparable performance with less than one-fourth of the original visual words. On the other hand, compared to the dense bi-scale description, *weighting* the power of visual words improves the classification in a statistically significant manner (p-value < 0.032 for $k \in [3, 5]$): it brings an additional gain of accuracy of 3.4 percentage points at $k = 5$ neighbors, with a resulting accuracy of 90.2% (93.7% sensitivity, 86.2% specificity).

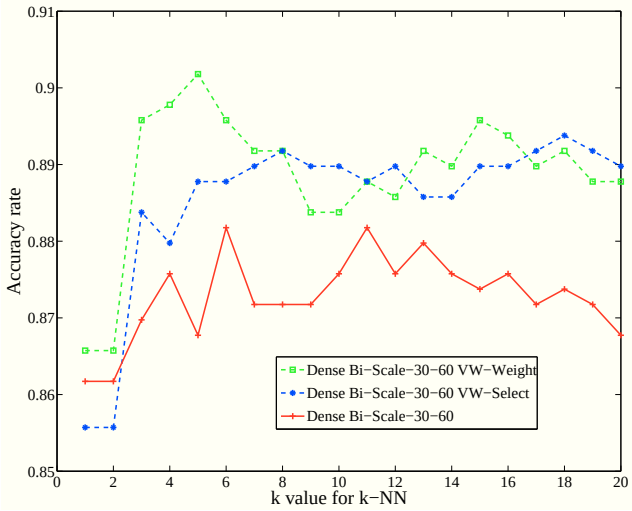


Figure 2.9: LOPO classification of pCLE mosaic images (with the default value $\theta = 0$) using the discriminative power of the visual words. The mosaic images have been built with non-rigid registration.

2.4.3 Statistics on Spatial Relationship between Local Features

Endoscopists establish their diagnosis on pCLE images from the examination of microscopic texture and shapes, but also of more macroscopic patterns. This suggests that the spatial organization of the goblet cells must be included in the retrieval process because it is essential to differentiate benign from neoplastic tissues. In the field of computer vision, several CBIR methods have been proposed that account for the spatial relationship between local image features. For example, Lazebnik et al. [Lazebnik 06] presented a spatial pyramid framework for the recognition of scene categories based on global geometric correspondence. More recently, Jegou et al. [Jegou 08] proposed to add a geometrical verification that takes spatial information into account. However, these methods are based on the assumption that they want to retrieve images of the exact same scene, which is not the case for our application.

Our objective in this section is to introduce a geometrical verification process after the retrieval process to remove possible retrieval outliers. A retrieval outlier should be defined as an image which is not visually similar to the query image. However, for this database, we do not have any quantitative measure of perceived similarity with respect to the query image. For this reason, we estimate outliers based on criteria that are complementary to the visual word signatures. In this study, outlier estimation is based on a supervised criterion that uses the most discriminative spatial relationships between visual features.

In order to introduce spatial information, we define the co-occurrence between two visual words using the natural 8-adjacency graph between the corresponding disk regions that compose the detection grid. Thus, we are able to store in a co-occurrence matrix M of size $K \times K$ the probability for each pair of visual words of

being adjacent to each other, as illustrated in Fig. 2.10. We investigated this idea in a prior study [André 09b]. Due to the symmetric property of M, its dimensionality is equal to $K(K+1)/2 = 5050$. By construction, the normalized co-occurrence matrix is a histogram, so the vector of its lower triangular elements defines a spatial signature. Then, one could use this spatial signature for a mosaic image, or its concatenation with the standard visual word signature. However, given the relatively small number of mosaic images, 499 exactly in our database, the 5050 elements of the spatial signature are too numerous to parameterize a mosaic image: using them for the retrieval would lead to over-fitting.

To focus on the discriminative information in the co-occurrence matrix but reduce its dimensionality, we chose to apply a linear discriminant analysis (LDA). Using the textual diagnostic information in the database, we aim at differentiating, in a supervised manner, the images of the benign class from the images of the pathological class. The lower triangular elements of the co-occurrence matrix M are stored in a l x 1 dimensional vector denoted by $\mathbf{m}$, where l is equal to the number of lower triangular elements. The LDA weights, represented as a l x 1 dimensional vector denoted by $\mathbf{L}$, satisfy:

$$\mathbf{L} = \mathbf{\Sigma^{-1}} (\mu_\mathbf{1} - \mu_\mathbf{2}) \tag{2.5}$$

where the l x l dimensional matrix Σ is the covariance matrix of the vector $\mathbf{m}$ associated with all training images, and where the l x 1 dimensional vector $\mu_\mathbf{i}$ is the mean of the vector $\mathbf{m}$ associated with all the training images belonging to the

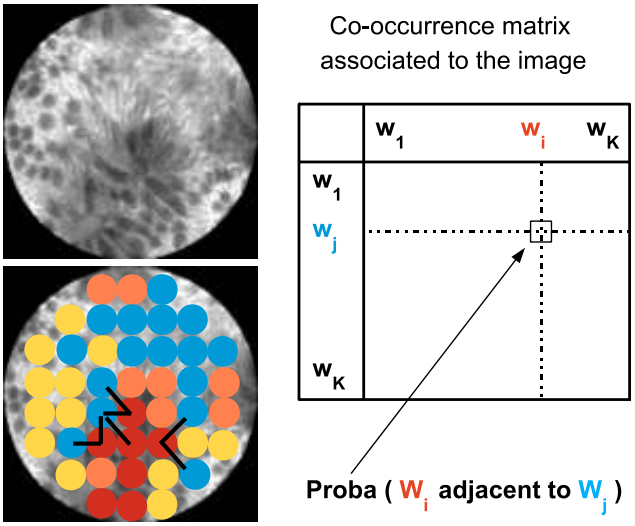


Figure 2.10: Example of a co-occurrence matrix M associated with a benign image. M is a symmetric matrix of size $K \times K$ where K is the number of visual words. Considering 2 visual words, respectively associated with the colors blue and red, black edges link the blue-labeled regions and the red-labeled regions that are adjacent to each other in the image. The number of these edges, after normalization, gives the probability that these 2 visual words are adjacent to each other in the image.

class C. Then, the most discriminative linear combination of the elements of $\mathbf{m}$ is the scalar value α which is given by the dot product: $\alpha = \mathbf{L}.\mathbf{m}$.

After the retrieval, outliers can be rejected during the verification process by thresholding on the absolute difference between the α value of the query and the α value of each retrieved image. Given a query image, every training image is a candidate neighbor of the query. Any training image which is estimated as an outlier with respect to the query is removed from the set of candidate neighbors. Then, the k nearest neighbors to the query are computed from the set of the remaining candidate neighbors, as shown in Fig. 2.11.

In practice, to prevent from over-fitting on our database, the number of LDA weights in the computation of the spatial criterion α had to be restricted. For this reason, we only performed a one-scale description and stored the $K = 100$ diagonal elements of the matrix M in the vector $\mathbf{m}$ for the LDA. The values of the threshold λ_α were chosen by analyzing the distribution of α across the benign and pathological images: $\lambda_\alpha = 2.6$ when considering only the disks of radius 60 pixels, and $\lambda_\alpha = 2.4$ when considering only the disks of radius 30 pixels. For the classification of mosaic images, Fig. 2.12 shows that, when added to the one-scale description with disks of radius 30 pixels, the outlier removal improves the classification accuracy, with statistical significance (p-value < 0.045 for $k \in [1, 4]$). At $k = 3$ neighbors, the corresponding gain of accuracy is 2.6 percentage points, with a resulting accuracy of 83.2% (82.8% sensitivity, 83.6% specificity). Besides, when added to the one-scale description but with disks of radius 60 pixels, the outlier removal brings an additional gain of accuracy. However, we established that this gain is not statistically significant (p-value ≥ 0.30 for $k \in [1, 10]$). This might be due to the size of our database: more information is captured at scale 60, so more data is needed to represent the variability of spatial relationships.

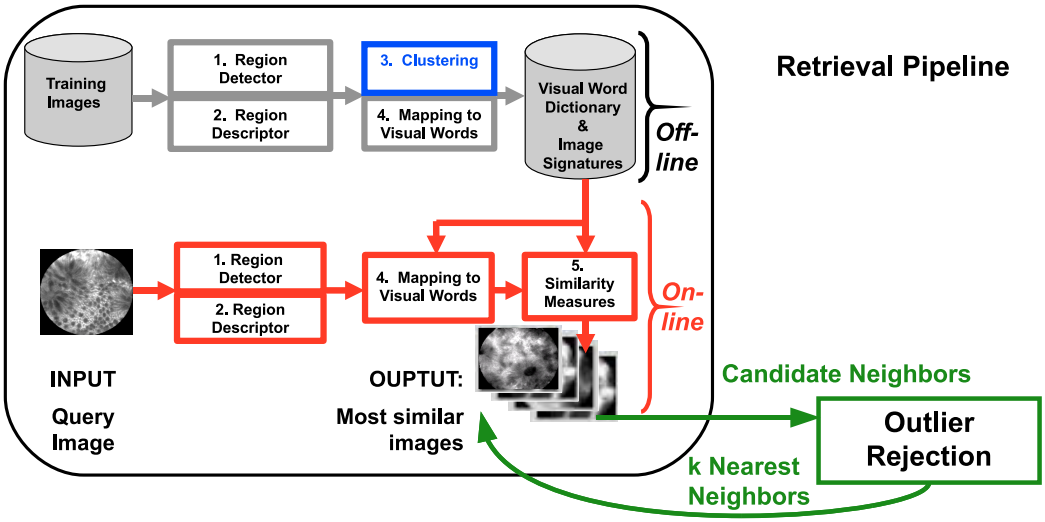


Figure 2.11: Overview of the retrieval pipeline followed by a the geometrical verification process performing outlier rejection.

In fact, the efficiency of our geometrical outlier removal method highly depends on the size and the representativity of the training database, which is still not large enough with respect to the high dimensionality of the co-occurrence matrix of visual words. More work is thus needed to better exploit the co-occurrence statistics. In particular, it would be relevant to use the correlatons proposed by Savarese et al. [Savarese 06], which are built from clustering correlogram elements. Indeed, correlograms are able to capture spatial co-occurrences of visual words at multiple scales, and their clustering is a way to solve the over-fitting issue by reducing the dimensionality.

The performance of our outlier removal process based on geometrical verification can also be qualitatively appreciated in Figs. 2.13, 2.14 and 2.15, showing some typical results as well as worst results.

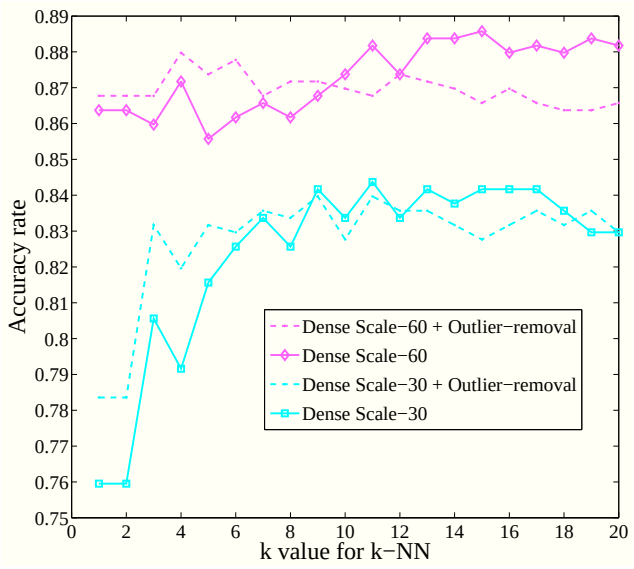


Figure 2.12: LOPO classification of pCLE mosaic images (with $\theta = 0$) using outlier removal. The mosaic images have been built with non-rigid registration.

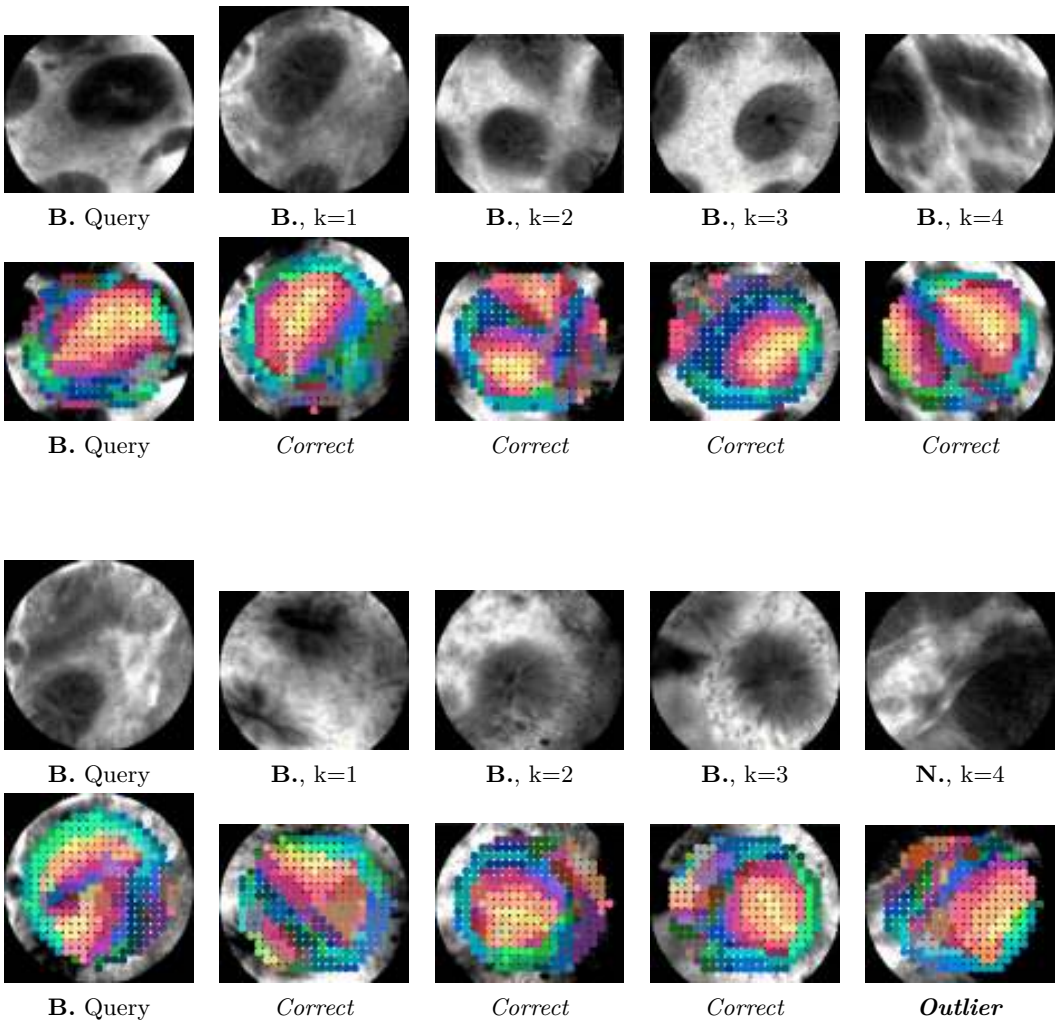


Figure 2.13: Typical image retrieval results provided by our method from two benign queries. **B.** indicates Benign and **N.** Neoplastic. From left to right on each row: the queried image, and its k-NNs on the top layer, and their respective colored visual words on the bottom layer. An outlier is indicated by ***Outlier*** if it has been rejected by the spatial verification process, and by ***Error*** otherwise. FoV of the images: 240 µm.

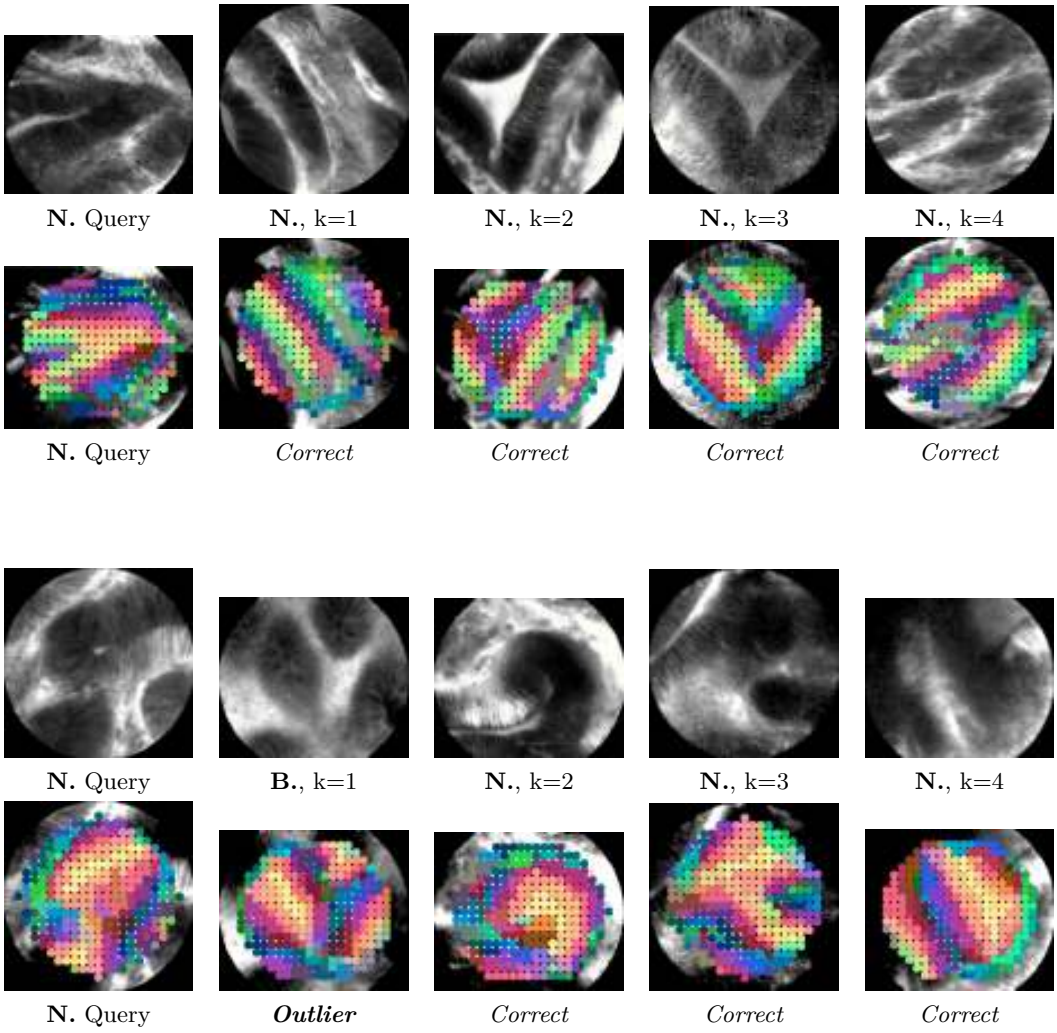


Figure 2.14: Typical image retrieval results provided by our method from two neoplastic queries. **B.** indicates Benign and **N.** Neoplastic. An outlier is indicated by ***Outlier*** if it has been rejected by the spatial verification process, and by ***Error*** otherwise.

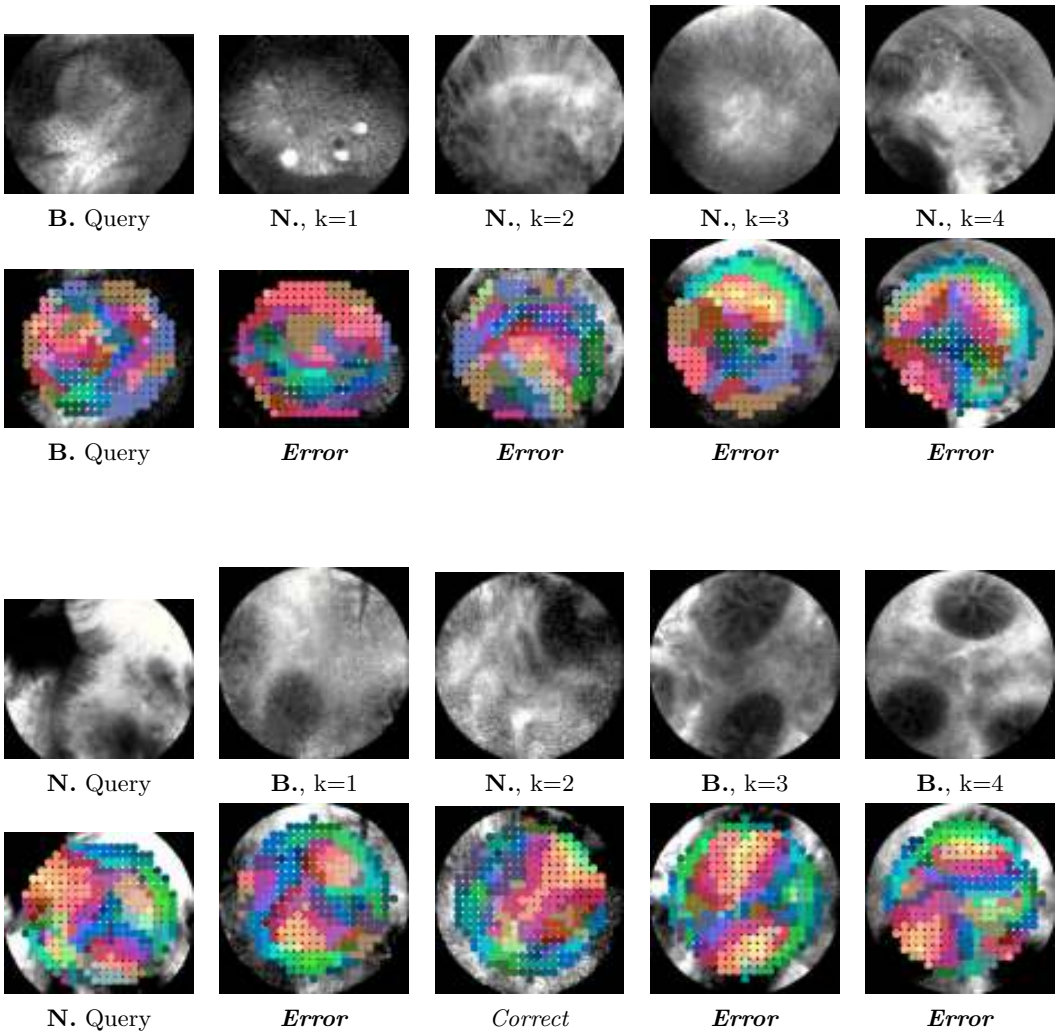


Figure 2.15: Worst image retrieval results provided by our method. The benign query on the top is a rare benign variety which is not represented in the training dataset. The neoplastic query on the bottom contains on its top left corner a partially visible elongated crypt which could not be totally described. **B.** indicates Benign and **N.** Neoplastic. An outlier is indicated by ***Outlier*** if it has been rejected by the spatial verification process, and by ***Error*** otherwise.

2.5 Endomicroscopic Videos Retrieval using Implicit Mosaics

2.5.1 From Mosaics to Videos

Although the retrieval of mosaic images instead of single images provided quite satisfying retrieval results, the non-rigid registration of the mosaicing process requires a long runtime. On average, the whole video-mosaicing process takes approximately 2 seconds per frame, which is incompatible with a routine clinical practice. Besides, the temporal information of videos, which is lost in the mosaic image representation, may be used by the endoscopists, who consider the videos as useful for real-time diagnosis. It would therefore be of interest to keep this information in our retrieval system.

For this reason, we investigated Content-Based Video Retrieval (CBVR) methods to retrieve similar videos instead of similar images. Our idea, which we previously explored in a preliminary study [André 10c], consists of including in the retrieval process the possible spatial overlap between the images from the same video sequence. For an efficient video retrieval, our objective is to build one short signature per video, which not only enables a reasonable memory space to store training data, but also considerably reduces the retrieval run-time. We looked at a more effective method which could only use the coarse registration results of mosaicing, i.e. the translation results between successive frames, that are computed in real-time [Vercauteren 08] during the image acquisition time. Another means of dealing with the large computational resources required by the complete mosaicing algorithm might be to rely on highly efficient implementations of the underlying registration algorithms. Graphical processing units (GPU) have for example been successfully applied for such purpose by Modat et al. [Modat 10]. More work is needed to see whether these implementations would allow for a real-time implementation of the complete mosaicing algorithm.

To reach our objective of efficient video retrieval, we first compute independently the visual words in all the images belonging to the database of video sub-sequences. Then, for each sub-sequence, we use the translation results to build a map of the overlap scores of all local regions belonging to the images of the sequence, as illustrated Fig. 2.16 on the right: for each region, the overlap score is the number of overlapping input images in the region. To define the signature H_S of a video sub-sequence S, we propose to take, for each image I of the sequence, the number τ of overlapping images in each densely detected region r of I, and to weight the contribution of r to the frequency of its visual word by $1/\tau$. Let i be an index of one of the K visual words. $i(.)$ is a function that associates a region r to the index of the visual word to which the region r is mapped. $\Gamma(.)$ is a second function that associates a region r to the number of overlapping images in this region. The visual word histogram of the video sub-sequence is then defined by:

$$H_S(i) = \frac{1}{Z} \sum_{I \in S} \sum_{r \in I} \frac{\delta(i(r), i)}{\Gamma(r)} \tag{2.6}$$

In this formula, δ is the Kronecker notation and Z is a normalization factor, introduced to normalize the visual word histogram. Z corresponds to the total number of physical regions in the overlapping area. More precisely:

$$Z = \sum_{i=1}^{K} (\sum_{I \in S} \sum_{r \in I} \frac{\delta(i(r), i)}{\Gamma(r)}) \tag{2.7}$$

From the video sub-sequence signatures, we define a full video signature by considering the normalized sum of the signatures of the constitutive sub-sequences of the video. Thanks to this histogram summation technique, the size of a video signature remains equal to the number of visual words, which reduces both retrieval runtime and training memory. We call our method the "Bag of Overlap-Weighted Visual Words" (BoWW) method.

For our experiments, we perform a one-scale dense SIFT description with a grid spacing of 20 pixels, a disk radius of 60 pixels and $K = 100$ visual words. Retrieval results of our BoWW method applied on pCLE sub-sequences can be qualitatively appreciated, for benign and neoplastic queries, in Figs. 2.17, 2.18, 2.19, 2.20, 2.21, 2.22, 2.23, 2.24, 2.25 and 2.26.

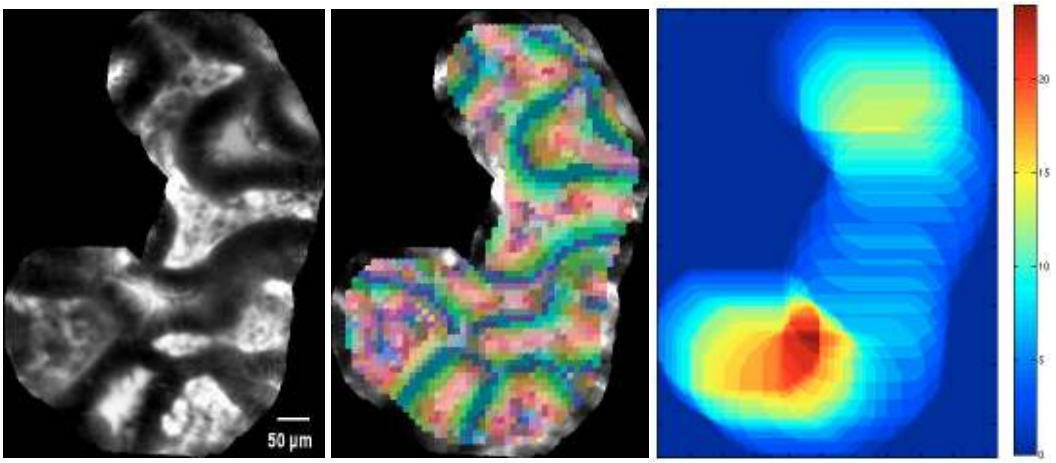


Figure 2.16: From left to right: Neoplastic pCLE mosaic obtained with non-rigid registration; Colored visual words mapped to the disk regions of radius 60 pixels in the mosaic image ; Overlap scores of the local regions in the mosaic space, according the translation results of mosaicing.

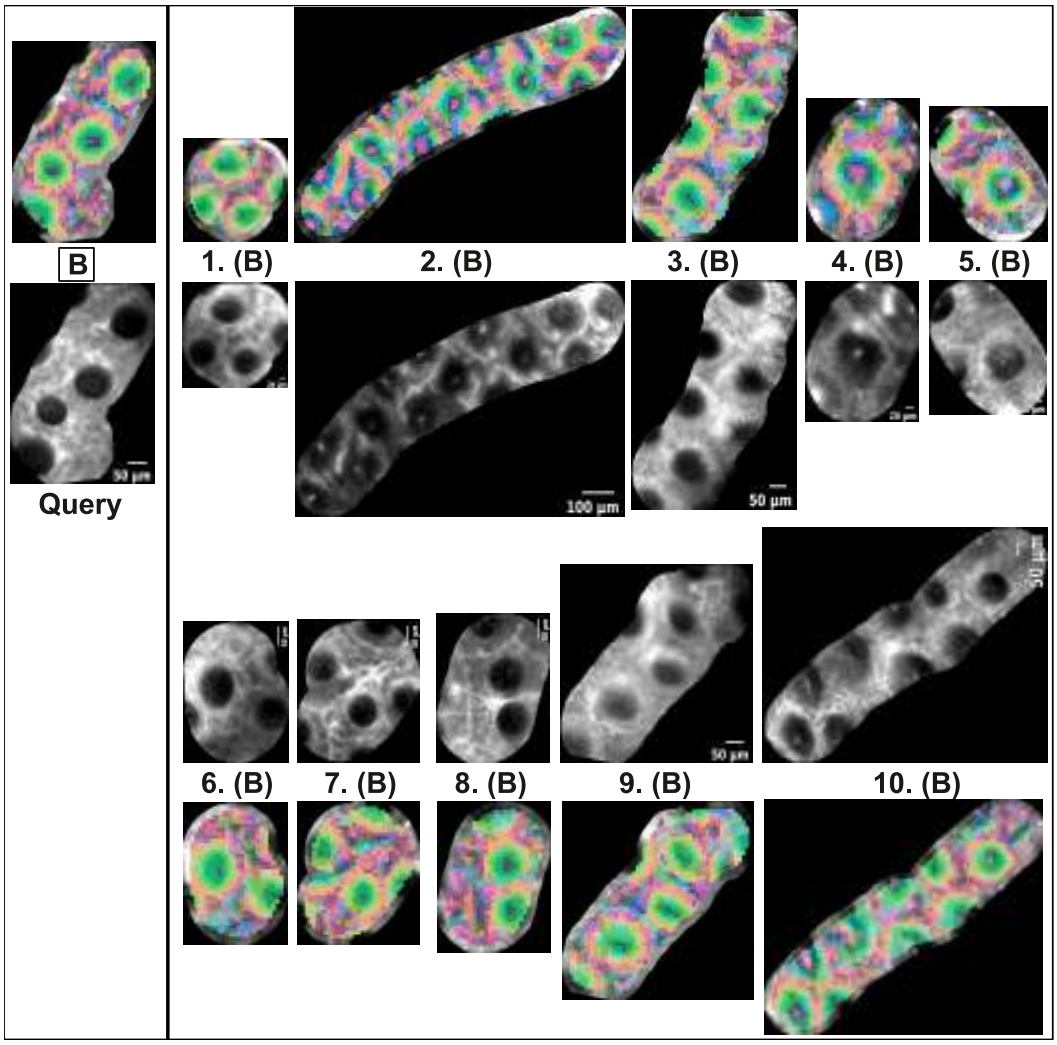


Figure 2.17: The 10 **most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method**. The pCLE video sub-sequences are represented by their corresponding fused mosaic image built with non-rigid registration, and are shown together with their visual words. **B** indicates Benign and **N** Neoplastic (not present here). For visualization purposes, the displayed visual words have been computed on the mosaic image on disks of radius 60 pixels. As a result, these colors are highlighting the geometrical structures in the mosaic images.

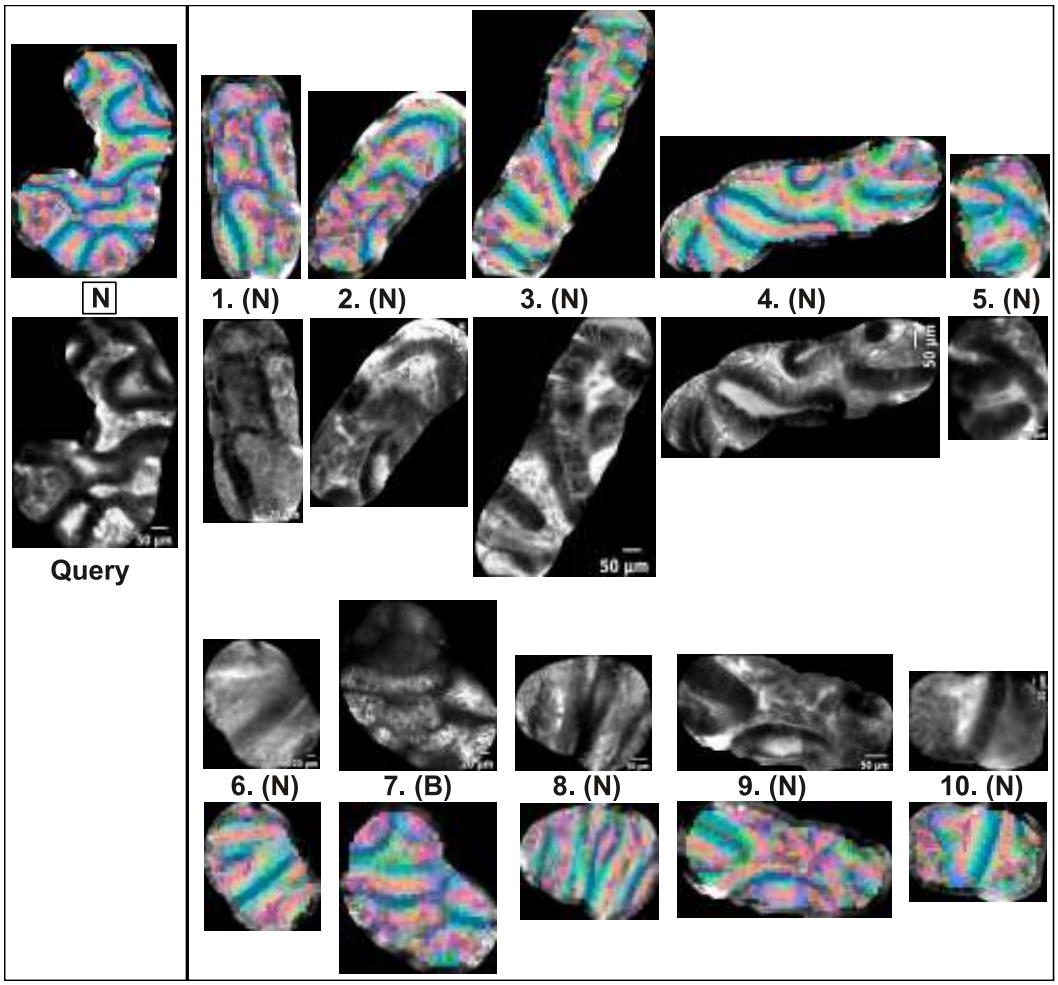


Figure 2.18: The 10 **most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign (not present here) and **N** Neoplastic

.

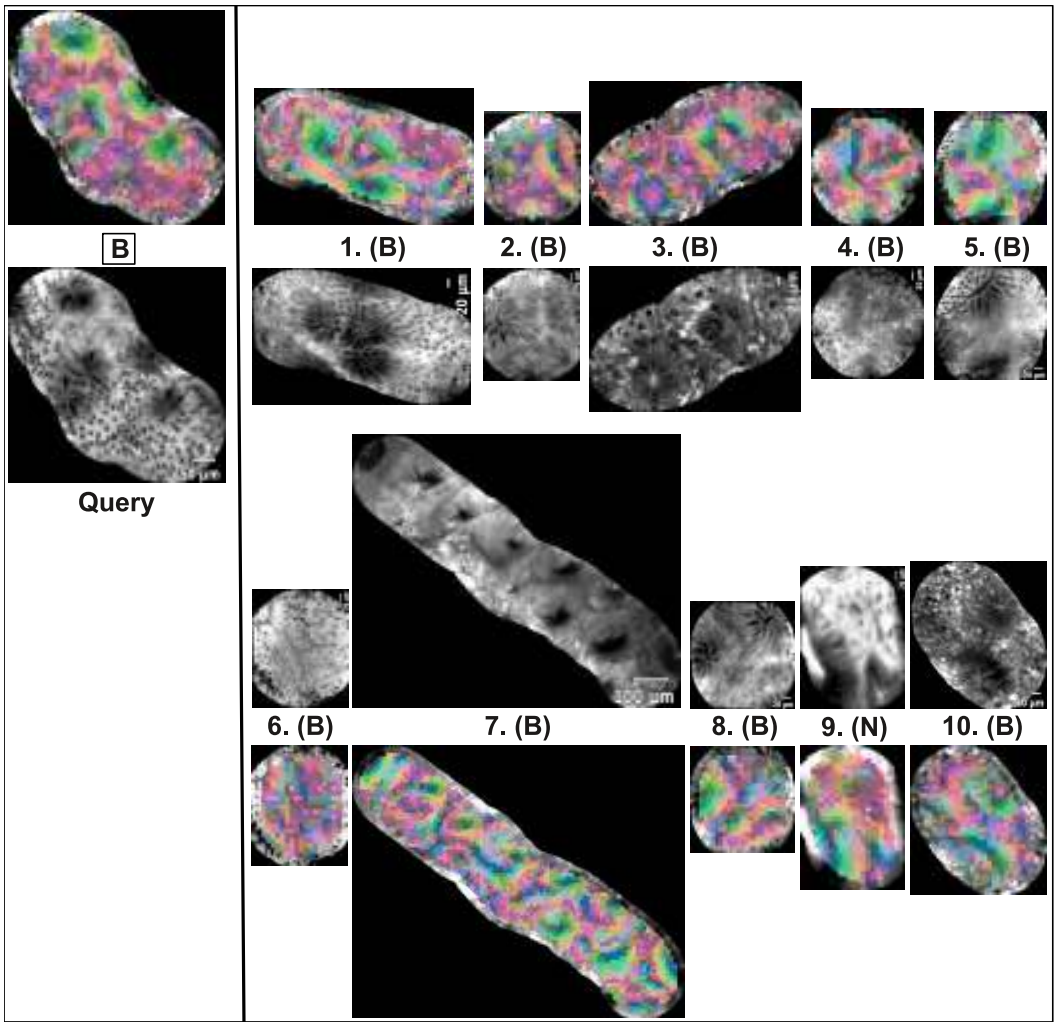


Figure 2.19: The 10 **most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method.** **B** indicates Benign and **N** Neoplastic.

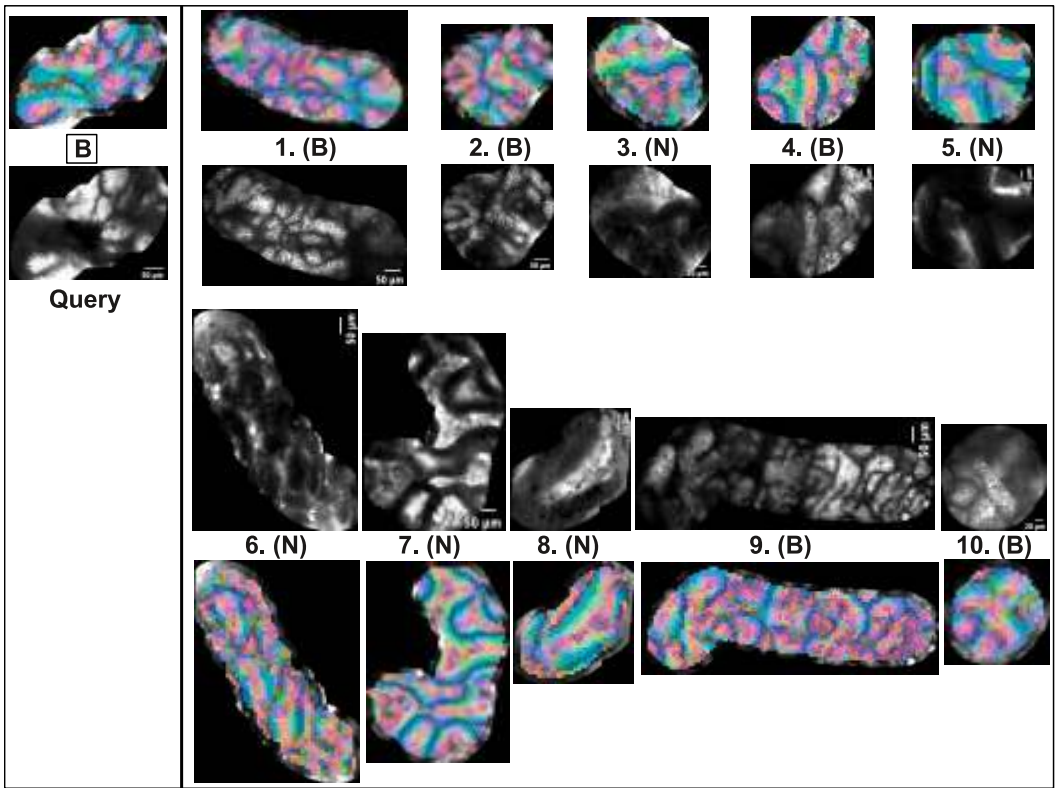


Figure 2.20: The 10 **most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method.** **B** indicates Benign and **N** Neoplastic.

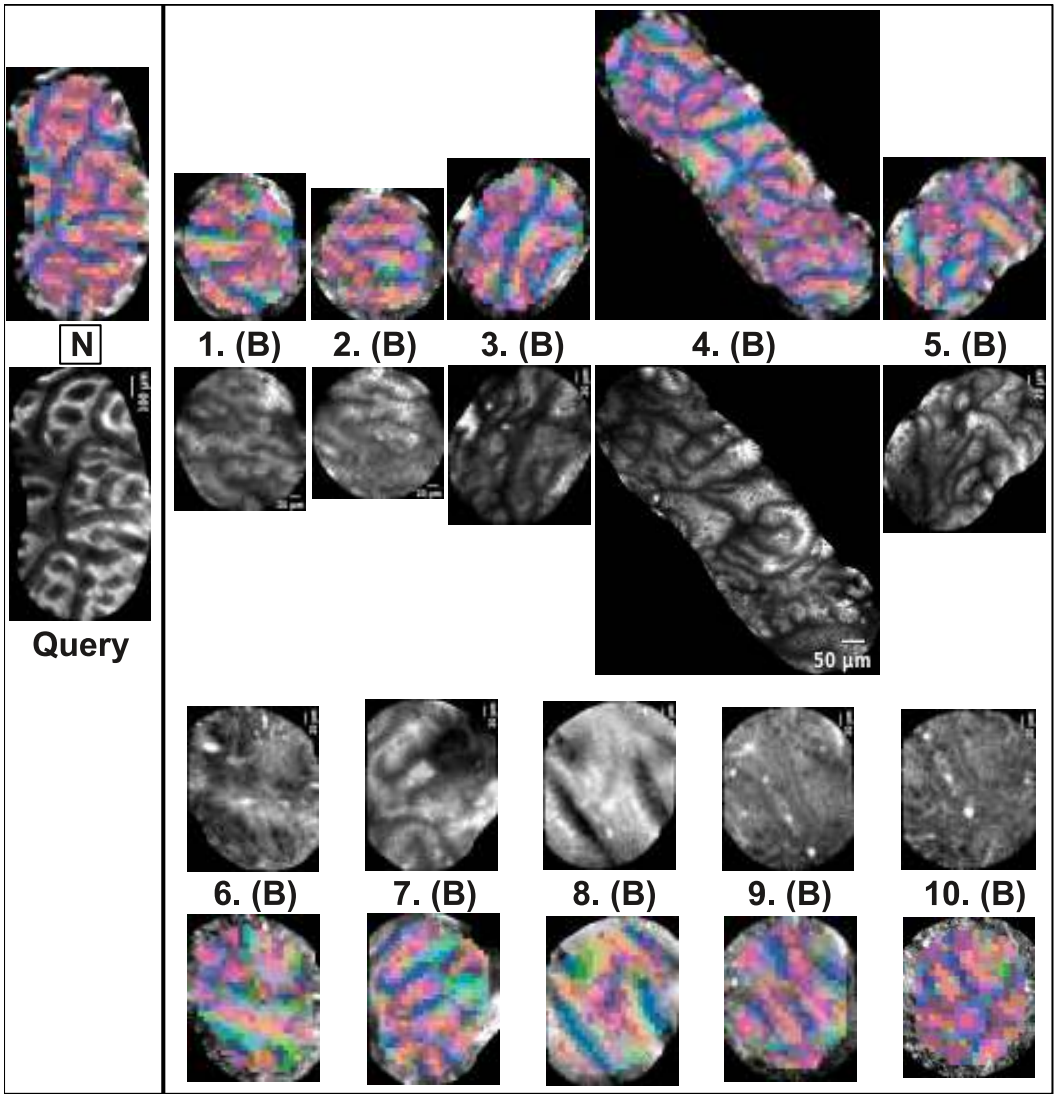


Figure 2.21: The 10 **most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign and **N** Neoplastic. This query is a rare variety of the neoplastic class. This is one of the worst retrieval results, that are due to the relatively small size and weak representativity of the training database.

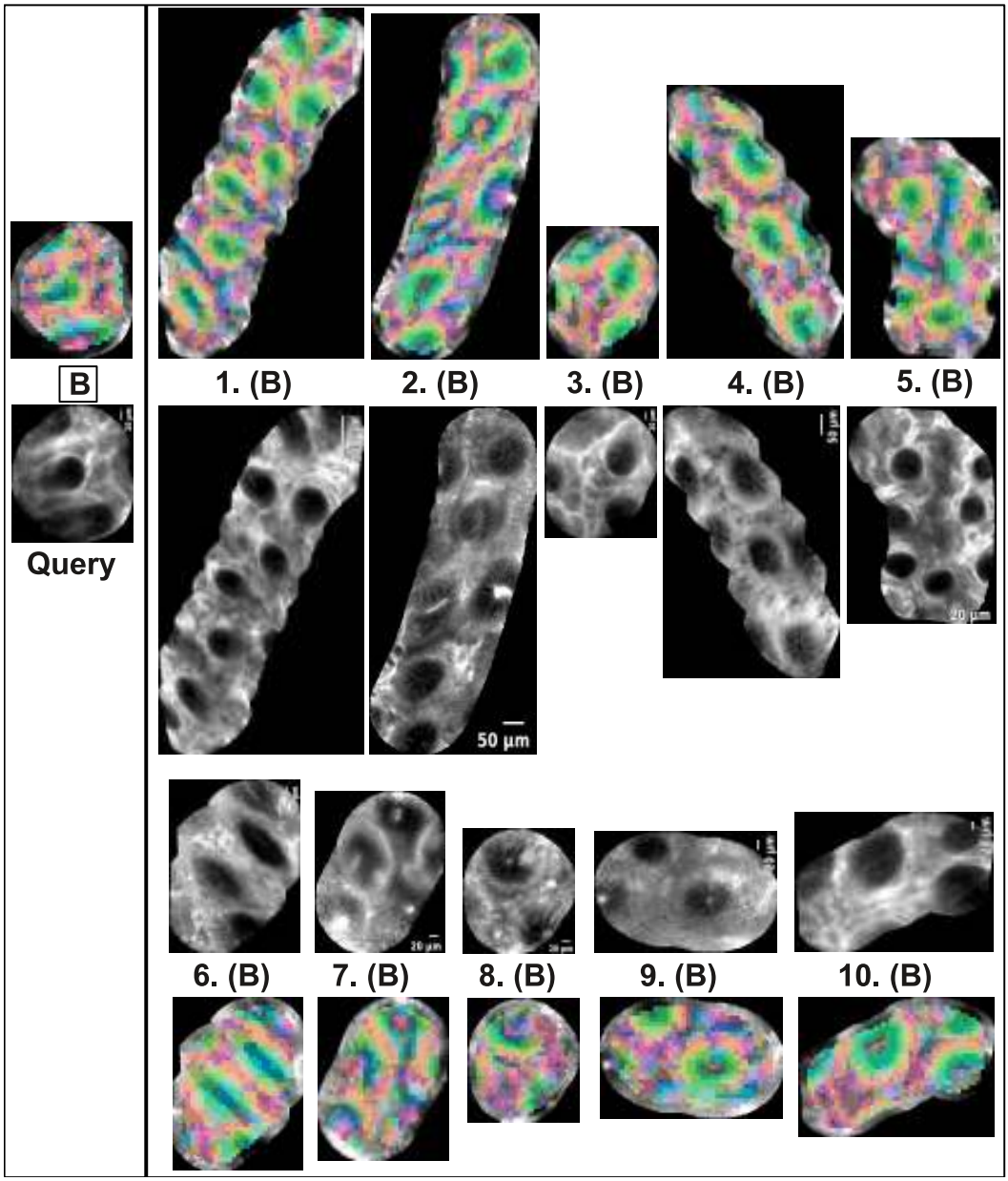


Figure 2.22: The 10 **most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method**. The pCLE video sub-sequences are represented by their corresponding fused mosaic image built with non-rigid registration. **B** indicates Benign and **N** Neoplastic (not present here).

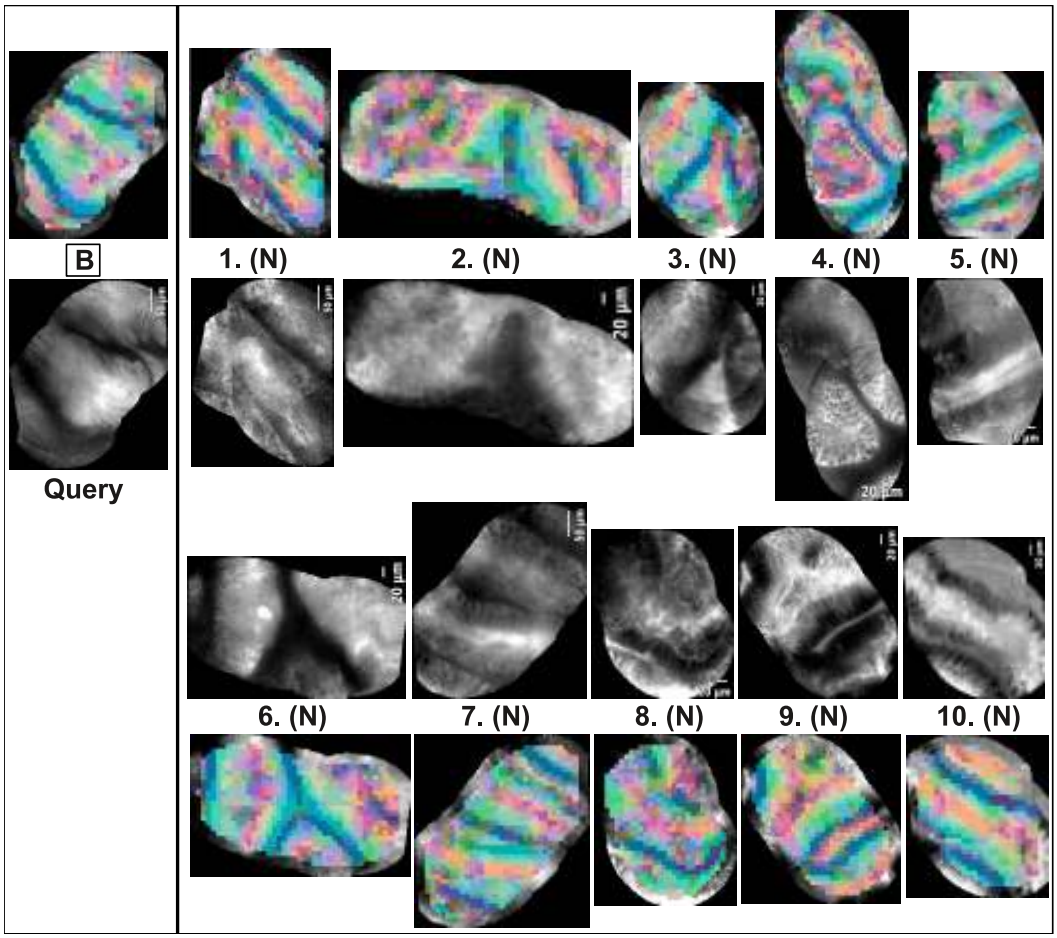


Figure 2.23: The 10 **most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign and **N** Neoplastic. Such bad retrieval result appears when the query is a rare variety of its pathological class, and is explained by the relatively small size and weak representativity of the training database.

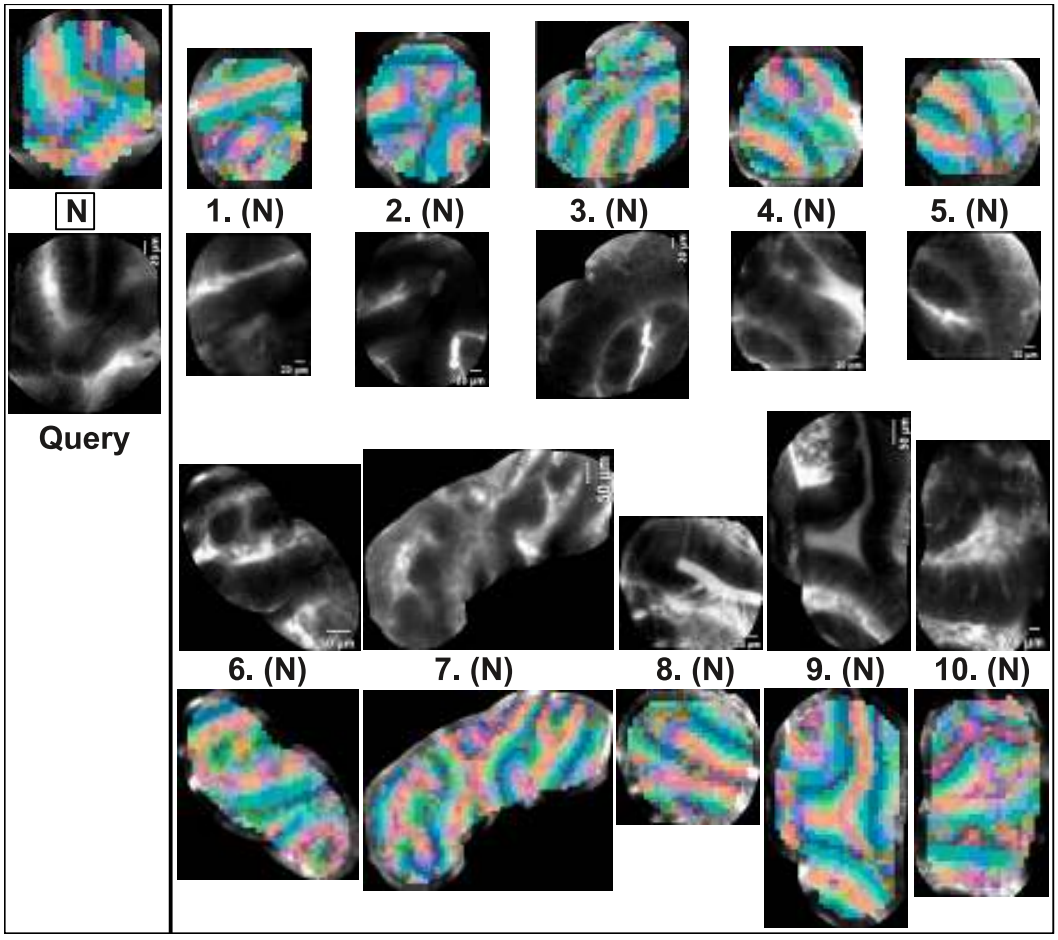


Figure 2.24: The 10 **most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign (not present here) and **N** Neoplastic.

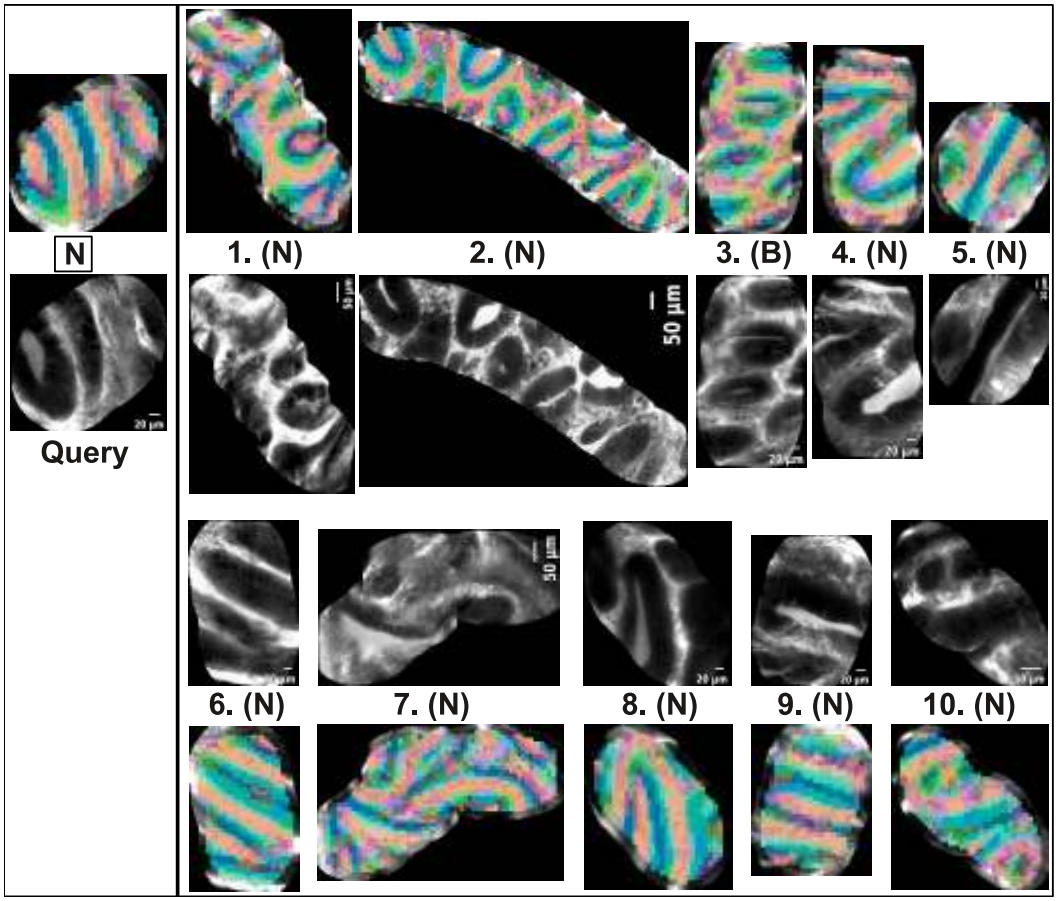


Figure 2.25: The 10 **most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method.** **B** indicates Benign and **N** Neoplastic.

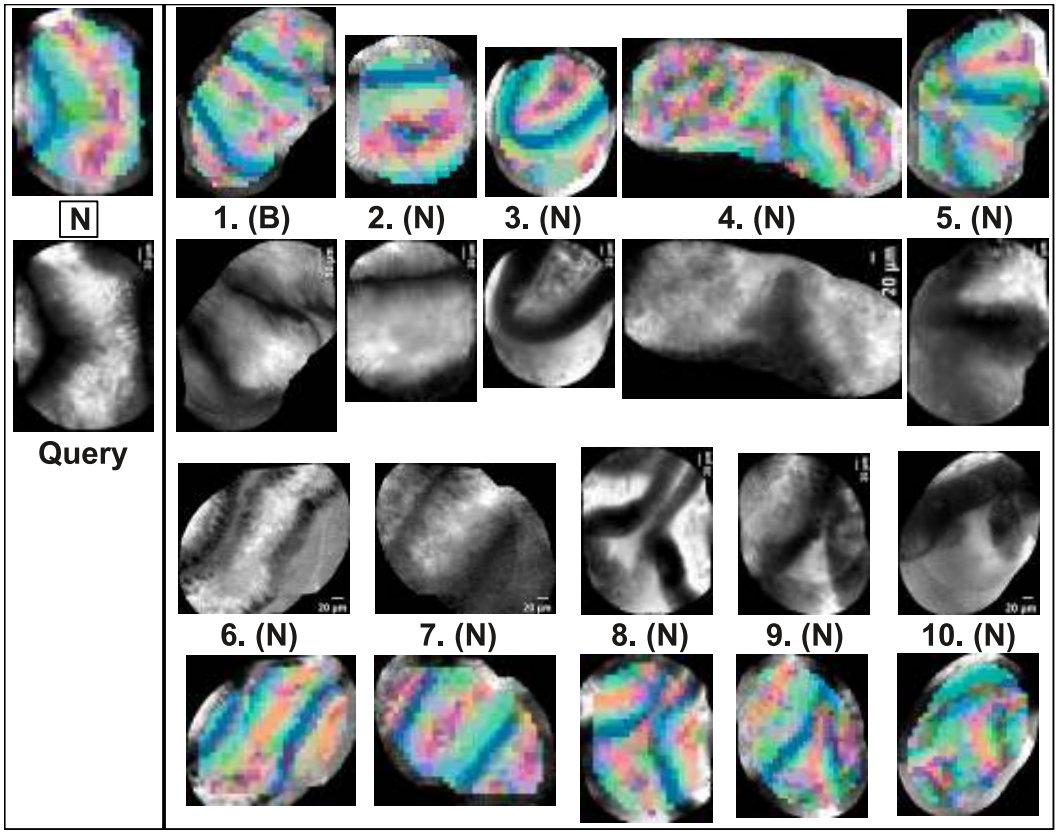


Figure 2.26: The 10 **most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign and **N** Neoplastic.

2.5.2 Method Comparison for Video Retrieval

In the previous sections, we have proposed different pCLE video retrieval techniques, such as overlap weighting and histogram summation, depending on the representation of the object of interest for the retrieval. The object of interest was either a video sub-sequence or a full video, and its representation was based on either single images or fused mosaic images. In order to evaluate these techniques, we define several methods that we will compare to each other. To establish statistical significance, the number of objects of interest that we classify needs to be sufficient to perform the McNemar's test, as explained in the Appendix A. This is always the case excepted for the 121 full videos for which statistical significance cannot be tested because the sum of differences is too small. A full video will either be considered as set of independent video sub-sequences or a set of independent single images. Then, each video sub-sequence will either be considered as a set of independent single images, a fused mosaic image, or an implicit mosaic made of the overlap-weighted single images.

For the classification of video sub-sequences, we call: "Weighted-ImOfMos" the method using the BoWW technique; "ImOfMos" the same method without overlap weighting ($\tau = 1$); "Mos" the method of Section 2.4.1 describing the single fused mosaic image obtained with non-rigid registration; and "AverageVote-Im" the method describing all the images independently and averaging their individual votes. For the classification of the full videos, the prefix "Sum-" means that we extended the methods with the signature summation technique to retrieve full videos as entities; "Sum-Im" is the method summing all the individual image signatures of the full video.

In this section, we also decide to compare the classification performances of our pCLE video retrieval methods with those of an efficient classification method: the NBNN classifier of Boiman et al. [Boiman 08], which was described in Section 2.2.4. Although NBNN classifies images, we can easily extend it to a "Weighted-NBNN" method for the classification of video sub-sequences, by weighting the closest distance computed for each region by the inverse of its overlap score. Then, by summing the accumulated distances, we can define the "Sum-Weighted-NBNN" method for the classification of full videos.

When comparing the methods for the classification of video sub-sequences, Fig. 2.28 shows that the accuracy of "Weighted-ImOfMos" is better than the accuracies of "AverageVote-Im" and "Weighted-NBNN", with statistical significance (p-value < 0.021 for $k \in [3, 10]$). For the classification of full videos, Fig. 2.29 shows that, from $k = 3$ neighbors, "Sum-Weighted-ImOfMos" has an accuracy which is better than the one of "Sum-Im", and equal or better than the one of "Sum-ImOfMos" and "Sum-Mos". The best full video classification result observed before 10 neighbors is achieved by "Sum-Weighted-ImOfMos" at $k = 9$, with an accuracy of 94.2% (sensitivity 97.7%, specificity 86.1%). At less neighbors, "Sum-Weighted-ImOfMos" already achieves a quite satisfying accuracy, e.g. 93.4% for 3 neighbors.

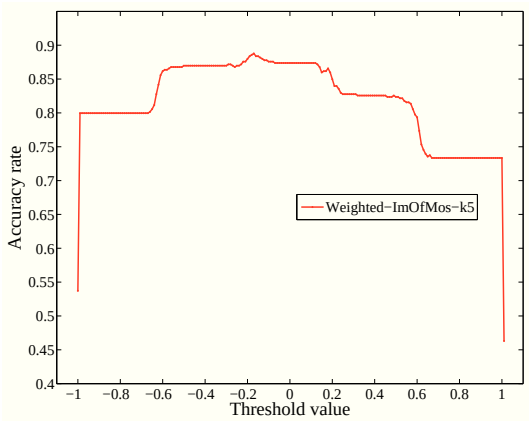


Figure 2.27: Accuracy rate for the classification of pCLE video sub-sequences by the LOPO Weighted-ImOfMos method, at $k = 5$ neighbors, depending on the value of the weighting threshold $\theta \in [-1, 1]$ that trades off the cost of false positives and false negatives. The slight accuracy peak at the negative value $\theta = -0.17$ reflects the fact that neoplastic features are more discriminative than the benign ones.

Besides, for each retrieval method and for a fixed number of neighbors, a peak of classification accuracy is reached at a θ value which is usually negative, as illustrated in Fig. 2.27 for the "Weighted-ImOfMos" method with a slight accuracy peak at $\theta = -0.17$ at $k = 5$ neighbors. This reflects the fact that neoplastic features are more discriminative than the benign ones, as described in Section 2.2.4. In fact, putting more weight on neoplastic patterns leads to increase the classification sensitivity, which is clinically important since it reduces the rate of false negatives.

Figs. 2.28 and 2.29 show that the ROC curves of the classification method "Weighted-NBNN" are not as good as the ROC curves of all the retrieval methods. with statistical significance for the classification of video sub-sequences (p-values ≤ 0.05). Besides, the best classification accuracies of video sub-sequences (resp. full videos) by "Weighted-NBNN" (resp. "Sum-Weighted-ImOfMos") are reached for $\theta_{\text{NBNN}} = 1.017 > 1$ (resp. $\theta_{\text{NBNN}} = 1.038 > 1$). This is also confirming that local neoplastic features are more discriminative than the benign ones, as described in Section 2.2.4.

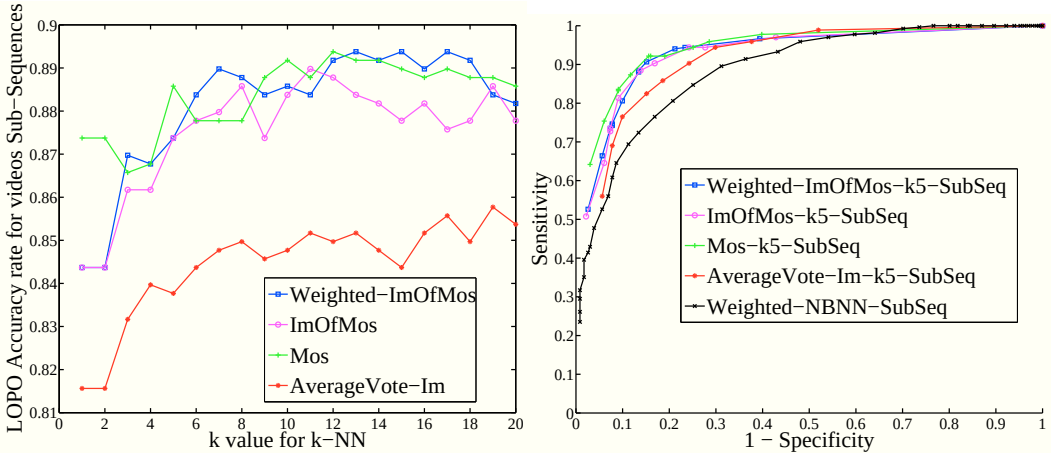


Figure 2.28: Left: LOPO classification of pCLE video sub-sequences, with the default value $\theta = 0$. The classification accuracy of "Weighted-NBNN" is 58.5% at the default value $\theta_{\text{NBNN}} = 1$, but it reaches 80.2% at the optimal value $\theta_{\text{NBNN}} = 1.017$. Right: Corresponding ROC curves at $k = 5$ neighbors with $\theta \in [-1, 1]$ and $\theta_{\text{NBNN}} \in [0, +\infty[$. θ and θ_{NBNN} trade off the cost of false positives and false negatives.

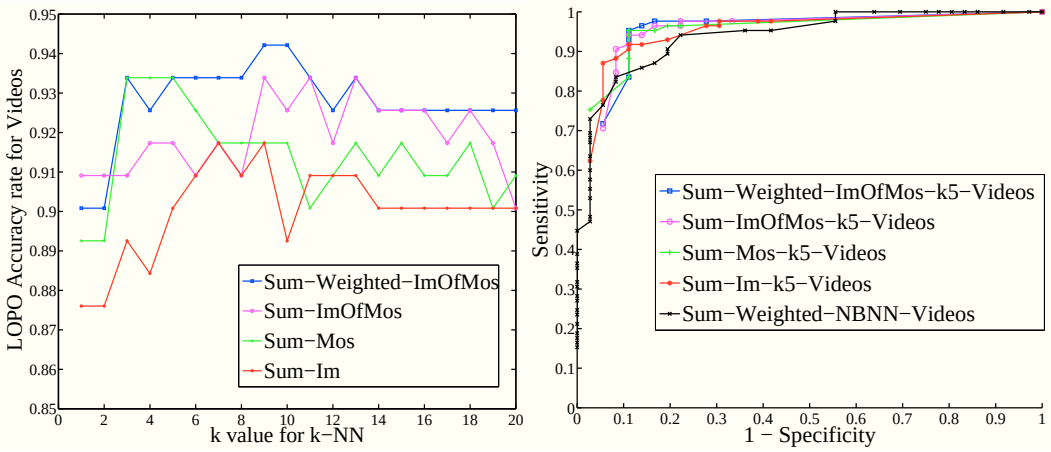


Figure 2.29: Left: LOPO classification of full pCLE videos, with the default value $\theta = 0$. The classification accuracy of "Sum-Weighted-NBNN" is 40.5% at the default value $\theta_{\text{NBNN}} = 1$, but it reaches 89.3% at the optimal value $\theta_{\text{NBNN}} = 1.038$. Right: Corresponding ROC curves at $k = 5$ neighbors with $\theta \in [-1, 1]$ and $\theta_{\text{NBNN}} \in [0, +\infty[$. θ and θ_{NBNN} trade off the cost of false positives and false negatives.

2.6 Finer Evaluation of the Retrieval

2.6.1 Diagnosis Ground Truth at a Finer Scale

In the previous sections, we used only two classes for retrieval evaluation because binary classification has a clinical meaning based on the distinction between neoplastic and non-neoplastic lesions, and thus delivers numbers that are easily interpretable by physicians. Nevertheless, in order to refine the quantitative evaluation of the retrieval, we decided to exploit diagnosis annotations available at a finer scale, and to perform a multi-class classification.

From the 121 videos of our database, 116 have been annotated at a finer scale by expert endoscopists, who define five subclasses to better characterize the colonic polyps. The benign class is subdivided into two classes: "purely benign lesion" (14 videos) and "hyperplastic lesion" (21 videos). The neoplastic class is subdivided into three classes: "tubular adenoma" (62 videos), "tubulovillous adenoma" (15 videos) and "adenocarcinoma" (4 videos).

2.6.2 Multi-Class Classification and Comparison with State of the Art

Based on the finer diagnosis ground truth, we perform a k-NN 5-class classification using LOPO cross-validation, and consider the overall classification accuracy (number of all correctly classified samples / total number of samples) as the evaluation criterion. For comparison with the state-of-the-art methods, the video sample size (116 annotated videos) is not sufficiently large to generate enough differences in the McNemar's test, as explained in the Appendix A. To be able to measure a statistical significance, we take as objects of interest mosaic images instead of videos, and we consider the 491 mosaics built from the 116 videos and we apply our Dense-Scale-60 method. The resulting evaluation of the methods for mosaic image retrieval using 5-class classification is shown in Figs. 2.30 and 2.31. Our annotated database is quite unbalanced with respect to the five subclasses, the most represented class ("tubular adenoma") being the pathology of highest prevalence. However, we checked that the naive classification method, which classifies all the queries in the class "tubular adenoma" and reaches an overall accuracy of 41.3%, is outperformed by the Dense-Scale-60 method from $k = 1$, and with statistical significance from $k = 3$. Although the overall accuracy of 56.8% reached by our method may appear low in terms of classification, it is a closer indicator of our retrieval performance. Moreover, we demonstrate that our mosaic retrieval method outperforms the state-of-the-art CBIR methods (Haralick, Texton, HH-SIFT) and the NBNN classifier, with statistical significance from 3 nearest neighbors.

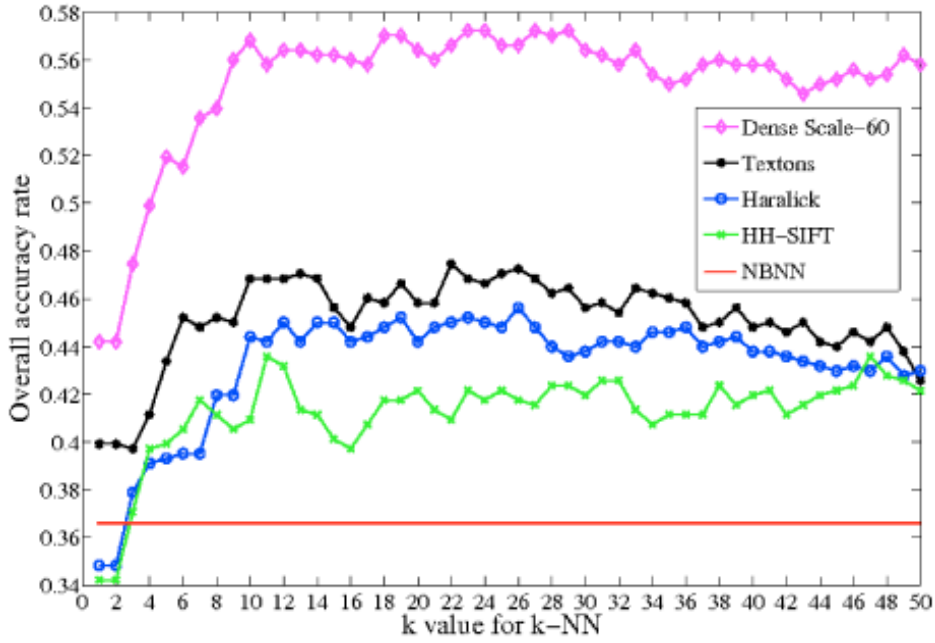


Figure 2.30: 5-class LOPO classification of pCLE mosaic images by the methods. The NBNN classification accuracy does not depend on k. The mosaic images have been built with non-rigid registration.

Method	Dense-Scale-60 $k = 10$	Textons $k = 10$	Haralick $k = 10$	HH-SIFT $k = 10$	NBNN
Accuracy	56.8 %	46.8 %	44.4 %	40.9 %	36.7 %
Statistical significance of Dense-Scale-60's gain		p-value < 0.0073 for $k \geq 3$	p-value < 0.0022 for $k \geq 1$	p-value < 0.00063 for $k \geq 1$	p-value < 0.0063 for $k \geq 1$

Figure 2.31: 5-class LOPO classification of pCLE mosaic images by the methods at k nearest neighbors. The statistical significance of the gain of the Dense-Scale-60 method is measured with the McNemar's test, as explained in the Appendix A.

2.7 Conclusion

To the best of our knowledge this study is the first approach to retrieve endomicroscopic image sequences by adapting a recent and powerful local image retrieval method, the Bag-of-Visual-Words method, introduced for recognition problems in computer vision.

By first designing a local image description at several scales and with the proper level of density and invariance, then by taking into account the spatio-temporal relationship between the local feature descriptors, the first retrieved endomicroscopic images are much more relevant. When compared to learning and retrieving images independently, our "Bag of Overlap-Weighted Visual Words" method using a video-mosaicing technique improves the results of video retrieval and classification in a statistically significant manner. With the vote of the $k = 9$ most similar videos, it reaches more than 94% of accuracy (sensitivity 97.7%, specificity 86.1%), which is clinically pertinent for our application. Moreover, fewer neighbors are necessary to classify the query at a given accuracy. This is relevant for the endoscopist, who will examine only a reasonably small number of videos, i.e. typically 3 to 5 similar videos. Besides, the video retrieval method is based on histogram summations that considerably reduce both retrieval runtime and training memory. This will allow us to provide physicians during ongoing endoscopy with whole annotated videos, similar to the video of interest. Such a pCLE video retrieval system potentially supports diagnostic decision and avoids unnecessary polypectomies of non-neoplastic lesions.

Besides, our generic framework could be reasonably applied to other organs or pathologies, and also extended to other image or video retrieval applications. Another clinical application would be the detection of neoplasia in patients with Barrett's esophagus, for which Pohl et al. [Pohl 08] already demonstrated the interest of endomicroscopy. Sharma et al. [Sharma 11] recently demonstrated that, when added to high-definition white-light endoscopy, pCLE significantly improved the ability to detect neoplasia in patients with Barrett's esophagus. Therefore, in Chapter 4, we will apply our video retrieval method to a pCLE database on the Barrett's esophagus.

Despite the lack of a direct objective ground truth for video retrieval, we evaluated our content-based retrieval method indirectly on a valuable database. By taking the k-NN classification accuracy as a surrogate indicator of the retrieval performance, we demonstrated that our retrieval method outperforms the state-of-the-art methods with statistical significance, on both binary and multi-class classification. Beyond classification-based evaluation, our goal in Chapter 5 will be to generate a perceptual similarity ground truth and directly evaluate the retrieval.

For future work, we plan to work on more complex description spaces, for example based on ellipsoidal regions, to better capture the elongated patterns in pCLE videos. We also plan to enlarge the training database. Indeed, a larger training database would not only improve the classification results if all the characteristics of the image classes are better represented, it would also allow us to exploit a larger

number of description attributes without facing the over-fitting issue. For example, the whole matrices of visual word co-occurrence at several scales could be better exploited. Potential ways of doing so include their incorporation into the description as proposed by Zhang et al. [Zhang 09], or their extraction at hierarchical scales in the image as described in the Hyperfeatures of Agarwal and Triggs [Agarwal 08]. On the other hand, the co-occurrence matrix could be better analyzed by more generic tools than Linear Discriminant Analysis. For example, a more complete spatial geometry between local features could be learned by considering the visual words as a Markov Random Fields model, whose parameters could be estimated using a method such as the one presented in [Descombes 99]. We also plan, for the testing process, to either use all the images of the tested video or to automate the splitting and the selection of video sub-sequences of interest. Besides, the learning process could leverage the textual information of the database. As for incorporating the temporal information, a more robust approach would not only consider the fused image of a mosaic but the $2D + t$ volume of the registered frames composing the mosaic. This would allow us to work on more accurate visual words and better combine spatial and temporal information. We could for example introduce spatio-temporal features, as those presented by Wang et al. [Wang 09], or as the 3-dimensional SIFT descriptor proposed by Scovanner et al. [Scovanner 07].

To conclude, the binary classification results that we obtained on our colonic polyp database compare favorably with the accuracy of pCLE diagnosis established on the same videos, among non-expert and expert endoscopists, for the differentiation between neoplastic and non-neoplastic lesions. Considering 11 non-expert endoscopists, the study of Buchner et al. [Buchner 09a] showed an interobserver agreement with an average accuracy of 72% (sensitivity 82%, specificity 53%). Considering 3 expert endoscopists, Gomez et al. [Gomez 10] obtained an average accuracy of 75% (sensitivity 76%, specificity 72%). The learning curve pattern of pCLE in predicting neoplastic lesions was demonstrated with improved accuracies in time as observers' experience increased. Thus, prospectively, our endomicroscopic video retrieval approach could be valuable not only for diagnosis support, but also for training support to improve the learning curve of the new endoscopists, and for knowledge discovery to better understand the biological evolution of epithelial cancers.

Chapter 3

A Clinical Application: Classification of Endomicroscopic Videos of Colonic Polyps

Table of Contents

Based on: [André 11a] B. André, T. Vercauteren, A. M. Buchner, M. Krishna, N. Ayache and M. B. Wallace. *Video retrieval software for automated classification of probe-based confocal laser endomicroscopy on colorectal polyps.* 2011. Article in submission. **Presented in the clinical abstract** [André 10b].

Introduction: *Whereas in the previous chapter we proposed a methodology for pCLE video retrieval, this chapter focuses on classification as a clinical application of our methodology. The objective of this chapter is to support* in vivo *diagnosis of*

colonic polyps, by designing a software for the binary classification between neoplastic and non-neoplastic lesions. We work on a database of pCLE videos of colonic polyps which is an extension of the pCLE database used in the previous chapter, because it includes new polyps for which the histological diagnosis is in contradiction with the pCLE diagnosis.

Methods: *Intravenous fluorescein pCLE imaging of colorectal lesions was performed on patients undergoing screening and surveillance colonoscopies, followed by polypectomies. All resected specimens were reviewed by a reference gastrointestinal pathologist blinded to pCLE information. Histopathology was used as gold standard for the differentiation between neoplastic and non-neoplastic lesions. The pCLE video sequences, recorded for each polyp, were analyzed offline by* 2 *expert endoscopists who were blinded to the endoscopic characteristics and histopathology. These pCLE videos, along with their histopathology diagnosis, were used to train the automated classification software which is a content-based retrieval technique followed by k-nearest neighbor classification. The performance of offline expert pCLE diagnosis was compared with that of automated pCLE classification. All evaluations were performed using leave-one-patient-out cross-validation.*

Results: 135 *colorectal lesions were imaged in* 71 *patients. Based on histopathology,* 93 *of these* 135 *lesions were diagnosed as neoplastic and* 42 *as benign. Compared to offline expert pCLE diagnosis, automated pCLE classification has statistically equivalent accuracy, sensitivity and specificity (respectively* 89.6%*,* 92.5% *and* 83.3%*). Moreover, the automated pCLE classification software provides, as intermediate results, several annotated videos that are visually similar to the pCLE video of interest and immediately tangible to the endoscopist.*

Discussion: *This study demonstrates that diagnostic performance of the automated method for classification of pCLE videos is high and comparable to the offline diagnostic performance of expert endoscopists. The automated pCLE classification software could thus help endoscopists in diagnosing pCLE videos online. In particular, it could be used as a second-reader tool to support pCLE diagnosis. Further studies are warranted to evaluate the impact of using the automated pCLE retrieval and classification software on the diagnostic performance of the endoscopists.*

French summary

Introduction : *Si dans le chapitre précédent nous avons proposé une méthodologie pour la reconnaissance de vidéos ECM, ce chapitre considère la classification comme une application clinique de notre méthodologie. L'objectif de ce chapitre est d'assister le diagnostic* in vivo *des polypes du côlon, en développant un système pour la classification binaire entre les lésions néoplasiques et non néoplasiques. Nous travaillons sur une base de données de vidéos ECM sur les polypes du côlon, qui est une extension de la base de données ECM utilisée dans le chapitre précédent. En effet,*

la nouvelle base inclut de nouveaux polypes pour lesquels le diagnostic histologique est en contradiction avec le diagnostic ECM.

Methodes : *L'imagerie ECM des lésions colorectales a été réalisée, après injection de fluorescéine intraveineuse, sur des patients qui ont eu une coloscopie de dépistage et de surveillance, suivie de polypectomies. Tous les spécimens réséqués ont été examinés par un médecin pathologiste de référence, en aveugle des informations ECM. L'histopathologie a été utilisée comme étalon pour la différenciation entre les lésions néoplasiques et non néoplasiques. Les séquences vidéo ECM enregistrées pour chaque polype ont été analysées par* 2 *endoscopistes experts, lors d'un examen hors-ligne, en aveugle des caractéristiques endoscopiques et histopathologiques. Ces vidéos ECM, ainsi que leur diagnostic histologique, ont été utilisés pour l'apprentissage de la classification automatique, qui repose sur une technique de reconnaissance par le contenu suivie d'une classification par plus proches voisins. La performance du diagnostic ECM établi hors-ligne par les endoscopistes experts a été comparée à celle de la classification ECM automatique. Toutes les évaluations ont été effectuées en utilisant la validation croisée de type "leave-one-patient-out cross-validation".*

Resultats : 135 *lésions colorectales ont été imagées dans* 71 *patients. D'après l'histopathologie,* 93 *lésions parmi les* 135 *lésions ont été diagnostiquées comme néoplasiques et* 42 *d'entre elles comme bénignes. En comparaison avec le diagnostic ECM établi hors-ligne par les endoscopistes experts, la classification ECM automatique a une précision, une sensibilité et une spécificité statistiquement équivalentes (respectivement* 89.6*%,* 92.5*% et* 83.3*%). D'autre part, le système de classification ECM automatique fournit, comme résultats intermédiaires, plusieurs vidéos ECM annotées qui sont visuellement similaires à la vidéo d'intérêt et immédiatement tangibles pour l'endoscopiste.*

Discussion : *Cette étude démontre que la performance diagnostique de la méthode de classification automatique des vidéos ECM est élevée et comparable à la performance du diagnostic ECM établi hors-ligne par les endoscopistes experts. Le système de classification ECM automatique pourrait ainsi aider les endoscopistes à diagnostiquer en ligne les vidéos ECM. En particulier, il pourrait être utilisé comme un outil de deuxième lecture pour assister le diagnostic ECM. Des études supplémentaires sont nécessaires pour évaluer l'impact de l'utilisation du système de reconnaissance et de classification ECM sur les performances diagnostiques de l'endoscopiste.*

3.1 Introduction

Colorectal cancer is the second leading cause of cancer-related death in the United States [Hawk 05]. Its development includes several morphological stages, from benign to adenomatous polyps with low grade dysplasia to adenocarcinoma. Suspicious lesions are usually detected with standard colonoscopy by the endoscopists who either perform confirmatory biopsy, or if high certainty exists, perform imme-

diate therapy such as resection or ablation of diseased tissue. Because standard endoscopic imaging can only diagnose disease states with moderate levels of certainty [Norfleet 88, Rastogi 09], histopathology remains the gold standard for final diagnosis [Winawer 06]. However, the requirement for *ex vivo* histology implies a large proportion of unnecessary polypectomies and often requires a separate endoscopic procedure to be performed for treatment. It also increases the cost of colorectal cancer screening.

Probe-based Confocal Laser Endomicroscpy (pCLE) enables the endoscopist to image the epithelial tissue *in vivo*, at the microscopic level with a confocal miniprobe, and in real-time during ongoing endoscopy. Preliminary findings by Meining et al. [Meining 07] demonstrated the applicability of pCLE in diagnosing colorectal neoplasia *in vivo* with high sensitivity and specificity (93% and 92% respectively) in 13 patients with colorectal lesions. In a recent study including a large pool of 75 patients, Buchner et al. [Buchner 10] compared offline pCLE diagnosis to virtual chromoendoscopy (NBI and FICE) and showed that offline pCLE had higher sensitivity (91% versus 77%) with similar specificity (76%). As noted by Wallace and Fockens [Wallace 09], the current challenge for the endoscopists is *in vivo* diagnosis using pCLE.

In order to provide an objective support for pCLE diagnosis, we aim at designing a computer-based system for the automated classification of colonic polyps into neoplastic and non-neoplastic lesions. For this application, a content-based image retrieval (CBIR) approach is relevant because, contrary to "black box" classification systems, a CBIR-based classification system extracts, from a training database, annotated pCLE videos that are visually similar to the video of interest and immediately tangible to the endoscopist. The pathology of the video query is estimated from the histopathological votes of these already diagnosed videos. Another advantage of CBIR-based classification is that the extracted similar videos can be presented to the endoscopist in a second reader paradigm to better support pCLE diagnosis. Thus, even though CBIR-based classification is not the most powerful way of performing classification, it offers a second reading of the pCLE data.

The main goal of this study is to compare, using the same database of colonic polyps, the clinical performances of our automated pCLE classification software with those of offline pCLE diagnosis by endoscopists expert in pCLE, with histopathology remaining the gold standard reference.

3.2 Patients and Materials

3.2.1 Patients

The patients included in the study were enrolled between November 2007 and March 2009 for previous studies approved by Mayo Clinic Institutional Review Board, and from which we collected all available data to ensure an as large as possible sample size. These patients were enrolled for the study of Buchner et al. [Buchner 10] and for further studies of the same Mayo Clinic group. Only the patients with

complete diagnostic data are considered in our study. All study participants gave full written consent. Patients were enrolled if they were due for surveillance or screening colonoscopies, evaluation of known or suspected polyps on other imaging modalities, and endoscopic mucosal resection of larger flat colorectal neoplasia. Exclusion criteria were patients with non corrected coagulopathy, women who were pregnant or breast feeding, documented allergy to fluorescein, and patients with no colorectal lesions found during a study colonoscopy. Twenty-four hours before the procedure, patients were prepped with $2 - 4L$ polyethylene glycol solution. Conscious sedation was performed with intravenous administration of midazolam and meperidine.

3.2.2 Endoscopy Equipment and Procedure

All procedures were performed by either Michael B. Wallace or Anna M. Buchner using a high-definition colonoscope (Fujinon EC450HL5 or 490 ZW, Fujinon, Ft Wayne, NJ; Olympus CFH180, Olympus, Center Valley, NY). The system was equipped with the EPX 4400 processor (Fujinon Inc) or CV 180 Exera (Olympus, Co). The primary screening method was white-light high-definition colonoscopy. Then, either FICE mode 4 with Fujinon colonoscope or NBI with Olympus 180 series scope was used to characterize lesions in all patients.

The surface pit pattern of the lesion was classified according to the Kudo classification [Kudo 96] which is presented in Fig. 3.1. Anatomical site and morphological class of lesions were recorded in accordance with the Paris classification [ParisWorkshop 03] which is shown in Fig. 3.2. Fluorescein sodium $2.5-5.0\ mL$ 10% (AK Fluor, Akorn Pharmaceutical, Lake Forest, IL) solution was administered intravenously after the first polyp was identified. Immediately after fluorescein injection, pCLE video sequences of the lesions were acquired and recorded. According to the visual examination of both endoscopic and pCLE images, real biopsies were targeted to the most suspicious parts of the polyp. Appropriate treatment procedures, ranging from simple polypectomies to complex endoscopic mucosal resection of lesions, were then performed.

3.2.3 pCLE Acquisition Protocol

During a pCLE acquisition protocol, the endoscopist inserts, through the working channel of a standard endoscope, a confocal miniprobe (Coloflex UHD, Cellvizio GI) of external diameter 2.5 mm, which is made of $30,000$ optical fibers bundled together. The pCLE imaging setup is shown in Fig. 3.3. As a result, pCLE images of field-of-view 240 μm are acquired and reconstructed at a rate of 9 to 12 frames per second. In stable pCLE video sequences the miniprobe is in constant contact with the tissue. Representative endoscopic, pCLE, and histopathology images of tubular adenoma are shown in Fig. 3.3.

Prior to pCLE evaluation of the study polyps, the 2 expert endoscopists (Michael B. Wallace, Anna M. Buchner) viewed extensive published material on pCLE and

performed a self-calibration on training pCLE videos of 20 polyps of known pathology (10 adenoma and 10 benign). These "training" polyps were evaluated by a gastrointestinal pathologist (Murli Krishna) and came from 9 patients not included in the study. Once acquired, the pCLE videos of the study lesions were evaluated offline and in random order by the 2 experts, who were blinded to histology diagnosis and endoscopic appearance of the lesion. Offline pCLE diagnosis was made based on the established modified Mainz criteria [Kiesslich 04] for diagnosis of colorectal neoplasia, according to pit pattern and overall crypt and vessel architecture. Of the whole pCLE video imaging a polyp, the sequence of the video containing the most malignant pCLE features was considered to represent the polyp.

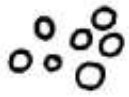
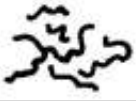
Pit type	Characteristics	Appearance using HMCC	Pit size (mm)
I	Normal round pits		0.07+/– 0.02 mm
II	Stella or papillary		0.09+/– 0.02 mm
IIIs	Tubular/round pits smaller than pit type I		0.03+/– 0.01 mm
IIIL	Tubular/large		0.22+/– 0.09 mm
IV	Sulcus/gyrus		0.93+/– 0.32 mm
V(a)	Irregular arrangement and sizes of IIIL, IIIs, IV type pit		N/A

Figure 3.1: Modified Kudo criteria. Colonic polyps are classified into pit pattern types according to their endoscopic appearance in "en-face" view, using HMCC (High Magnification Chromoscopic Colonoscopy). Type I and II are designated as non-neoplastic patterns whereas the other types are designated as neoplastic. Figure taken from [Hurlstone 08].

3.2.4 Histopathology as Gold Standard Diagnosis

All resected specimens were reviewed by a reference gastrointestinal pathologist (Murli Krishna) blinded to the pCLE information. Only the size and anatomic location were provided, which is the routine clinical practice at the Mayo Clinic institution. Intraepithelial neoplasia was defined using modified Vienna criteria

Endoscopic appearance	Paris class		Description
Protruded lesions	Ip		Pedunculated polyps
	Ips		Subpedunculated polyps
	Is		Sessile polyps
Flat elevated lesions	0–IIa		Flat elevation of mucosa
	0–IIa/c		Flat elevation with central depression
Flat lesions	0–IIb		Flat mucosal change
	0–IIc		Mucosal depression
	0–IIc/IIa		Mucosal depression with raised edge

Figure 3.2: The Paris classification. Colonic polyps are classified into Paris classes according to their histological appearance in transverse view. Figure taken from [Hurlstone 08].

[Schlemper 00, Rubio 06]: hyperplastic polyps were classified as benign lesions, while tubular adenoma, villous adenoma, tubulovillous adenoma and adenocarcinoma were classified as neoplastic lesions.

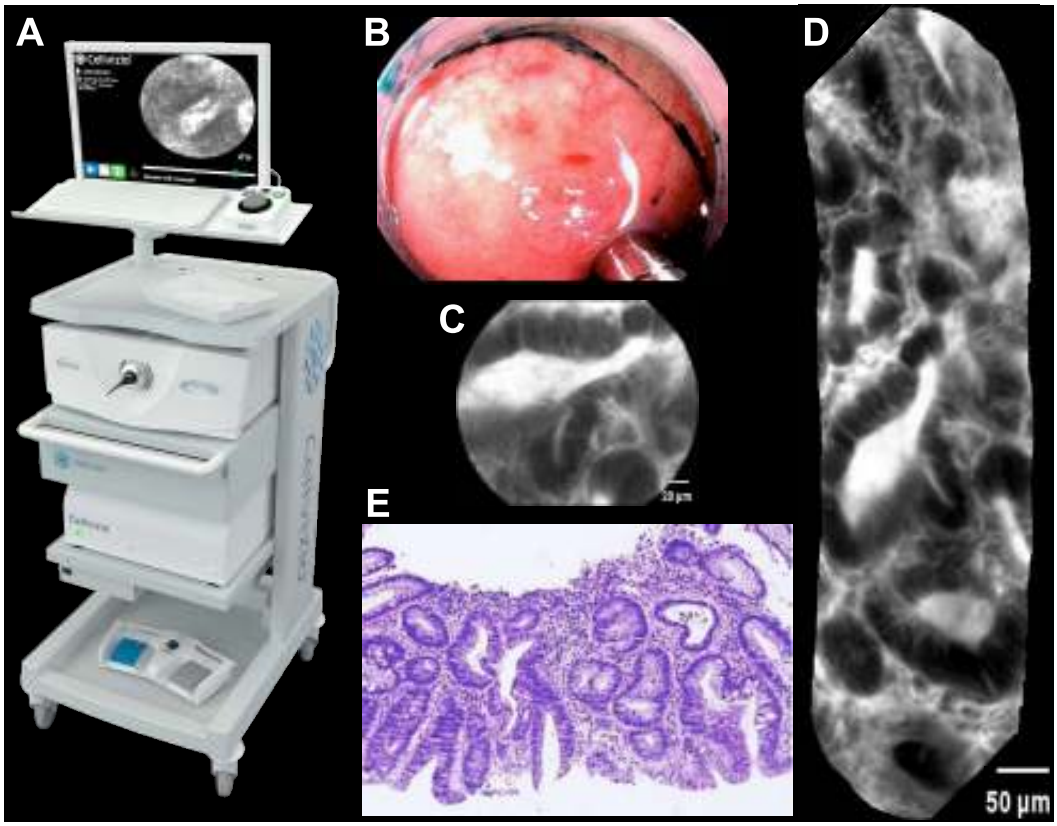


Figure 3.3: **(A)** Setup of pCLE imaging system (Cellvizio, Mauna Kea Technologies). **(B)** Endoscopic image of tubular adenoma, and the pCLE miniprobe. **(C)** An image of the pCLE video sequence. **(D)** A pCLE mosaic image built with the video mosaicing tool. **(E)** Histopathology image.

3.3 Methods

This section provides for the physicians a brief description of the methodology presented in Chapter 2 for pCLE video retrieval.

As the endoscopists use perceptual similarities between pCLE videos of known diagnosis to establish a diagnosis on a new pCLE video, we propose a content-based retrieval approach to design the automated pCLE video classification method. We revisited the standard Bag-of-Visual-Words (BoW) technique which has been successfully used in many content-based image retrieval applications in computer vision [Zhang 07].

3.3.1 Standard BoW Technique for Content-Based Image Retrieval

Standard BoW technique for image retrieval can be decomposed into four steps: region detection on the image, description of the regions, discretization of the feature space and similarity measuring between images. The detection step extracts salient regions in the image using sparse detectors. During the description step, a descriptor computes for each salient region its description vector. Then, the discretization step uses the result of a clustering method that builds K clusters, i.e. K visual words, from the union of the description vector sets gathered across all the images of the training database. Each description vector counts for one visual word, so an image can be represented by a signature of size K which is the histogram of its visual words. By construction, image signatures are invariant by viewpoint changes (image translation, rotation and scaling) and affine illumination changes. Finally, the similarity measuring step defines the similarity distance between two images as the as an adequate distance between their signatures: the most similar training images to the image of interest are defined as being the closest ones in terms of this distance.

3.3.2 Adjusting BoW Technique for pCLE Video Retrieval

First, we observed that discriminative information is densely distributed in pCLE images. Second, we noticed that several pCLE image patterns have the same shape but represent different objects characterized by their different size (e.g. mesoscopic crypts and microscopic goblet cells both have a rounded shape). So pCLE image description must not be invariant by scaling. To avoid scale invariance and to extract all the image information, we decided to apply, instead of standard sparse detectors, a dense detector that is made of overlapping disks having a fixed radius of 60 pixels and localized on a dense regular grid every 20 pixels. We maintained the invariance by in-plane translation and rotation, because the pCLE miniprobe translates and rotates along the tissue surface. Besides, as the diffusion rate of fluorescein administered before imaging procedure decreases through time, invariance by affine illumination changes was also preserved.

Expert endoscopists pointed out that the field of view of single still images may not be large enough to make a robust diagnosis. So we decided to retrieve

not single images but complete videos, by using the video mosaicing technique presented in [Vercauteren 06, Becker 07] and available in the Cellvizio software, to account for spatial overlap between time-related images. Examples of mosaics built with the video mosaicing tool are shown in Figs. 3.3, 3.5 and 3.6. To ensure online retrieval, we used the translation results of the real-time version of the video-mosaicing technique to weight the contribution of each local image region to its visual word. Then, we computed the video signatures with a histogram summation technique. Fig. 3.4 presents the whole pipeline of our retrieval-based classification framework, which can be run online during ongoing colonoscopy.

3.3.3 Classification of pCLE Videos using Similarity Distance

Once the visual signature of the video query was computed, the k-Nearest Neighbor (k-NN) search step identified the k closest training videos to the video query, by relying on the similarity distance between the video signatures. We then used the known histopathology diagnosis of these training videos to classify the query video, either as neoplastic or as non-neoplastic. Each of the k most similar training videos

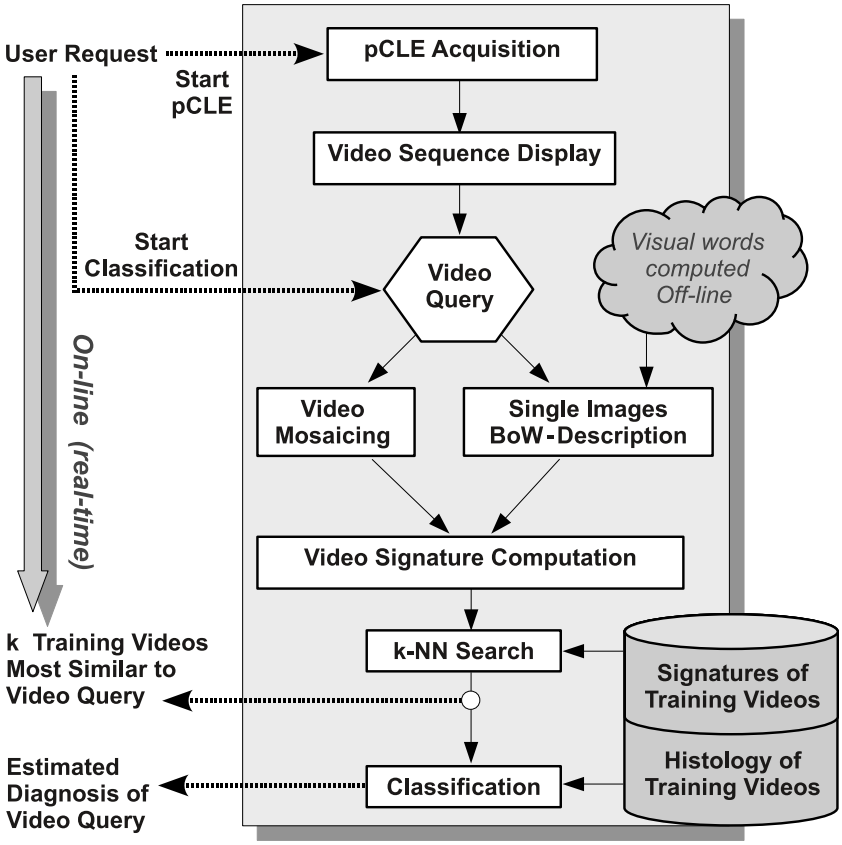


Figure 3.4: Pipeline of the pCLE retrieval-based classification framework, from the acquisition of the pCLE video query by the Cellvizio system to the online automated diagnosis estimation.

delivered a "histopathological" vote which is weighted by the inverse of its similarity distance to the video query.

Given the relatively small size of our pCLE database, we needed to learn from as much data as possible. We thus used the same database both for training and testing with cross-validation to avoid bias. As there were several videos acquired on the same patient, we performed a leave-one-patient-out cross-validation [Dundar 04]: all videos from a given patient are excluded from the training set before being tested as queries of our retrieval and classification methods. This also allowed us to find the optimal number of nearest neighbors, $k = 9$, which is the one that maximizes the accuracy of the retrieval-based classification results.

3.3.4 Statistical Analysis

To test for statistical difference between the two methods of interest, namely automated classification and offline classification by experts, we used McNemar's tests [Sheskin 11] and show the corresponding power calculations with a type I error $\alpha = 0.05$. Two-sided p-values < 0.05 were assumed to indicate statistical significance.

In order to assess statistical equivalence between the two methods, we used the two-sided Z-test between proportions [Jones 96] and computed 95% confidence intervals.

We refer the reader to the Appendix A for a detailed description of the McNemar's test and on the two-sided Z-test between proportions. Because the 135 pCLE videos constitute a small sample size, we used a correction for continuity for the McNemar's test.

The statistics on overall accuracy are dependent on the relative fraction of benign and neoplastic lesions examined, which in this study are 31.1% and 68.9%, respectively. Even though observations were made for more than one polyps in some patients, for the purposes of statistical analysis individual polyps (and their corresponding videos) were assumed to constitute independent observations. It is recognized that there was multiple testing of outcome data arising from individual polyps. Since the statistical tests were meant to highlight differences and since correction by Bonferroni's method would not have affected statistical significance in any of the comparisons, all p-values are presented uncorrected for multiple testing.

3.4 Results

3.4.1 Study Population and Colorectal Lesion Characteristics

Table 3.1 summarizes the demographic and general characteristics of the study population. None of the 71 patients experienced any endoscopic complications or adverse reaction to sodium fluorescein, with the exception of transient yellow discoloration of the skin and urine, which resolved by the time of discharge from the

recovery room (skin) or within 24 hours (urine). Histopathology and morphological classification of the 135 analyzed colorectal lesions are also provided in Table 3.1.

3.4.2 Qualitative Results: Visual Similarities between pCLE Videos

The pCLE database contains 135 pCLE videos representing each of the 135 polyps. The pCLE appearance of neoplastic lesions, compared to benign and hyperplastic lesions, included dilated irregular vessels, fluorescein leakage, cellular features of epithelial mucin depletion, and histological features of villiform crypts with increased optical density along epithelial border.

As the automated pCLE classification method is a similarity-based system that classified pCLE videos based on the votes of visually similar videos, its clinical relevance can be qualitatively evaluated by examining the intermediate results of

Study Population	**Summary ($n = 71$ patients)**
Age, median (min, $q1$, $q3$, max)	$75(46, 68, 79, 93)$
Gender, %	
Male	49
Female	51
History of colon cancer, %	9
Family history of colon cancer, %	10
Colorectal Lesions	**Summary ($n = 135$ lesions)**
Polyp size (mm), median (min, $q1$, $q3$, max)	$8(1, 5, 20, 60)$
Polyp location, %	
Cecum	24
Rectum	20
Ascending	18
Sigmoid	14.5
Transverse	15
Descending	5.5
Splenic flex	3
Histopathology diagnosis, %	
Hyperplastic	31
Tubular adenoma	52
Tubulovillous adenoma	11.5
Hyperplastic and adenomatous features	2.5
Adenocarcinoma	3
Neoplastic lesion, simplified histopathology, %	69
Paris classification, %	
1p	1
1s	57
2a	32
2b	5
2c	1
2a/c	4

Table 3.1: Study Population and Colorectal Lesions Characteristics $q1$ and $q3$ indicate respectively the first and the third quartiles.

video retrieval. Fig. 3.5 shows 4 typical results of the automated pCLE retrieval software. We observe that, despite the high variability in appearance of a given histopathological class (neoplastic or non-neoplastic), the automatically retrieved videos called "neighbors" look quite similar to the video queries, respectively $Q1$, $Q2$, $Q3$ and $Q4$. Besides, we notice that the closer the neighbor is to the query, the more similar it is to it.

In terms of classification, the pathological class is estimated by the weighted votes of the 3 retrieved neighbors. Video queries $Q1$, $Q2$, $Q3$ and $Q4$ have been correctly classified with respect to histopathology, both by automated classification and by expert endoscopists.

Fig. 3.6 shows 3 other results that reveal some limitations of the automated pCLE retrieval software. Video query $Q5$ corresponds to a rare variety of hyperplastic polyp correctly classified as non-neoplastic by the experts, but misclassified by the automated classification because it is not represented in the training database for retrieval. Video query $Q6$ corresponds to the ambiguous serrated adenoma case, correctly classified as non-neoplastic by the automated classification, but misclassified by the experts who consider serrated adenomas as malignant. Video query $Q7$ corresponds to a tubulovillous adenoma misclassified as non-neoplastic both by the experts and by the automated classification (this may be explained if a sampling error occurred and the corresponding biopsy was not performed exactly on the imaging spot).

3.4.3 Quantitative Results: Comparison with Expert Endoscopists

Classification accuracy, sensitivity and specificity of the two methods, automated pCLE classification (first method) and offline pCLE diagnosis of 2 experts (second method), are listed in Table 3.2. Automated classification reached a sensitivity of 92.5%, a specificity of 83.3% for a resulting accuracy of 89.6%. Expert review reached a sensitivity of 91.4%, a specificity of 85.7% and the same accuracy of 89.6%.

When testing for statistical difference, the p-values provided by McNemar's tests show that the differences between the 2 methods are not statistically significant (p-values > 0.05), and that there is very low power ($< 6\%$) to detect the observed differences.

When testing for statistical equivalence, the 95% confidence intervals provided by two-sided Z-tests between proportions are: -0.073 to 0.073 for the accuracy, -0.068 to 0.089 for the sensitivity and -0.18 to 0.13 for the specificity. These intervals include zero and are sufficiently small to suggest that the methods are equivalent. In particular, the -0.18 lower bound for the specificity is acceptable if the automated pCLE classification software is only taken as a second-reader tool to support pCLE diagnosis.

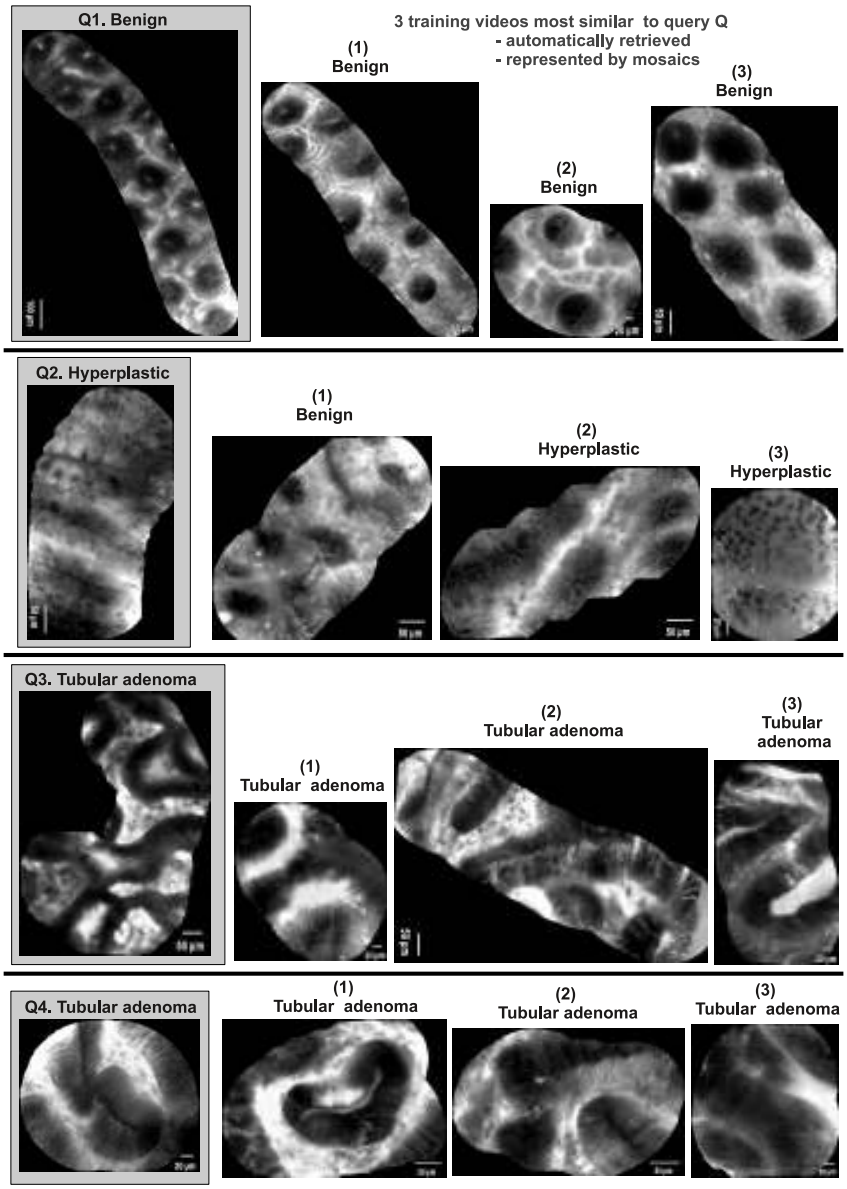


Figure 3.5: Typical results of automated pCLE video retrieval. The pCLE videos are represented by mosaic images; they are annotated with their histopathology diagnosis. Video queries are highlighted in gray and followed by their 3 most similar videos. Automated classification (hyperplastic versus neoplastic) of query videos is based on the votes of the similar videos. With respect to histopathology, both the automated classification and the pCLE diagnosis by experts are correct for these queries.

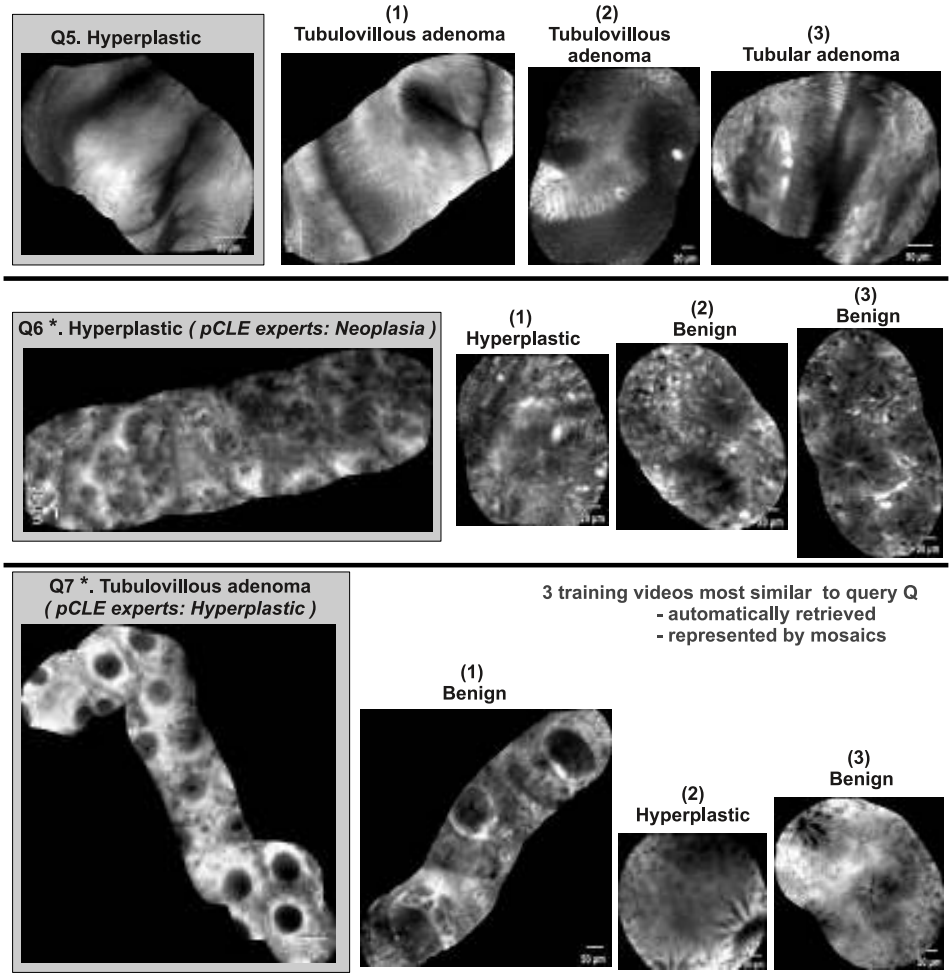


Figure 3.6: Results of automated pCLE video retrieval represented as mosaics. With respect to histology: the automated classification is correct for video query $Q6$ but incorrect for video queries $Q5$ and $Q7$, whereas the offline pCLE diagnosis by experts is correct for video queries $Q5$ but incorrect for video queries $Q6$ and $Q7$ (for which this disagreement is marked by $*$).

	(1) Automated Retrieval-based pCLE Classification	**(2) Offline pCLE Diagnosis of 2 Expert Endoscopists**
Accuracy		
%	89.6	89.6
Fraction	121/135	121/135
Sensitivity		
%	92.5	91.4
Fraction	86/93	85/93
Specificity		
%	83.3	85.7
Fraction	35/42	36/42
Statistical significance between (1) and (2) McNemar's test, $\alpha = 0.05$		
for Accuracy: (p-value, power)	(1, 2.5%)	
for Sensitivity: (p-value, power)	(0.82, 6.5%)	
for Specificity: (p-value, power)	(0.87, 5.2%)	
Statistical equivalence between (1) and (2) Two-sided Z-test		
95% CI for Accuracy	-0.073 to 0.073	
95% CI for Sensitivity	-0.068 to 0.089	
95% CI for Specificity	-0.18 to 0.13	

Table 3.2: Comparison of Accuracy, Sensitivity and Specificity between Automated Retrieval-based pCLE Classification and Offline pCLE Diagnosis by 2 Expert Endoscopists. CI stands for Confidence Interval.

3.5 Discussion

The present study demonstrates that, using a fairly representative database of colonic polyps, our automated method for the pCLE video classification had overall high accuracy, sensitivity and specificity, that are comparable to those of the offline pCLE diagnosis established by two endoscopists expert in pCLE. As the automated classification software can be run online during ongoing colonoscopy, it could be used as a second-reader tool to support and improve not only offline but also online pCLE diagnosis of endoscopists with various levels of expertise. In the majority of cases the second reader would agree with a moderately experienced endoscopist, who would be thus comforted in his/her diagnosis. For cases when they disagree, the endoscopist would have the opportunity to rethink his/her diagnosis and have more accurate *in vivo* interpretation. Besides, especially for small polyps, this second-reader tool could assist the endoscopist in adopting the "Diagnose, Resect and Discard Strategy" that dispenses with histopathological examination.

Gomez et al. [Gomez 10] analyzed *in vivo* pCLE interpretation in distinguishing between neoplastic and non-neoplastic lesions among 3 expert endoscopists and estimated an average accuracy of 75% (sensitivity 76%, specificity 72%) with good to moderate interobserver agreement. Buchner et al. [Buchner 09a] demonstrated that accurate interpretation of pCLE images by 11 endoscopists, considered as non expert in pCLE, can be learned rapidly with a short 2 hour training session. The learning curve pattern of pCLE in predicting neoplastic lesions was demonstrated with improved accuracies in time from 63% to 86% as observers' experience increased. Thus, prospectively, the automated classification method could be valuable not only for *in vivo* diagnosis support, but also for training support to improve the learning curve of the new endoscopists.

One of the advantages of our computer-based classification method is that it is not a "black box" but an informative tool based on the query by example model: it produces, as intermediate results, visually similar annotated videos that are immediately tangible to the endoscopist. From the qualitative observations of visual similarities between pCLE videos, we infer that the visually convincing results of the intermediate video retrieval step account for the relevance of the whole pCLE classification software. As few similar videos (less than 10) are necessary to classify a video query with a high accuracy, this visual information should be clinically useful for the endoscopist.

Further limitations of the classification software may include three main issues. First, a large training database is needed to be sufficiently representative of non-typical pCLE cases. This is even more challenging since the practice of pCLE is evolving and that new cases with atypical pCLE features may be still encountered. Second, the definition of "gold standard" for colorectal cancer screening is debatable because expert endoscopists and pathologists do not always agree. This could be illustrated by many examples of hyperplastic polyps redefined later as sessile serrated lesions by gastrointestinal pathologists, as in the study of Khalid et al. [Khalid 09]. The third limitation is that an obtained biopsy may be acquired unintentionally

from the area that does not correspond with the obtained pCLE imaging.

The task of the automated pCLE classification method is not to replace the endoscopist nor the pathologist but to assist the endoscopist in taking an informed decision. Before using the classification tool during an ongoing endoscopy procedure, more work is needed to improve its accuracy and to develop underlying tools that are both ergonomic and complementary. In particular, the online display of the retrieval outputs, for instance of the 3 most similar videos to the video query, together with their histopathology and possible multimodal clinical data, may be a precious underlying indicator for diagnosis decision. Such a sophisticated "Smart Atlas" for pCLE would allow the endoscopists in different centers to share and enrich their pCLE knowledge during ongoing endoscopy. Further studies are warranted to evaluate the impact of using automated pCLE retrieval and classification software on the pCLE learning curve and diagnostic performance of the endoscopists.

Chapter 4

Estimating Diagnosis Difficulty based on Endomicroscopy Retrieval of Colonic Polyps and Barrett's Esophagus

Table of Contents

Based on: [André 10a] B. André, T. Vercauteren, A. M. Buchner, M. W. Shahid, M. B. Wallace and N. Ayache. *An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis.* In Proceedings of the 13th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'10), pages 480–487, 2010. **Presented in the clinical abstract** [André 11b] with selected video abstract `http://www.youtube.com/watch?v=RVy-0Bxx9EQ`.

Learning medical image interpretation is an evolutive process that requires modular training systems, from non-expert to expert users. Our study aims at developing such a system for endomicroscopy diagnosis. It uses a difficulty *predictor to try and shorten the physician learning curve. As the understanding of video diagnosis relies on similarity-based reasoning, we propose a content-based video retrieval approach to estimate the level of interpretation* difficulty*. In addition to the pCLE database on colonic polyps used in the previous chapters, we introduce in this chapter a new pCLE database, on the Barrett's esophagus, to show the genericity of our retrieval method. Typical pCLE mosaic images of the Barrett's esophagus are illustrated*

in Fig. 4.1. The retrieval performance is evaluated indirectly using binary classification between pathological and non pathological cases. As shown in Chapter 2 on the Colonic Polyp *database, we demonstrate that, on the* Barrett's Esophagus *database, our retrieval method outperforms several state of the art methods. From our retrieval results, we then learn a* difficulty *predictor against a ground truth given by the percentage of false diagnoses among several physicians. Our experiments show that, for the two different databases, there is a significant correlation between our retrieval-based* difficulty *estimation and the* difficulty *experienced by the physicians.*

French summary

Apprendre à interpréter des images médicales est un processus évolutif qui nécessite des systèmes d'auto-formation modulaires, pour des utilisateurs non experts à experts. Notre étude vise à développer un tel système d'auto-formation pour le diagnostic en endomicroscopie. Ce système utilise un prédicteur de difficulté *dans le but de raccourcir la courbe d'apprentissage des médecins. Comme la compréhension du diagnostic établi sur des vidéos repose sur un raisonnement par similarité, nous proposons d'utiliser les résultats de la reconnaissance de vidéos par le contenu afin d'estimer la* difficulté *d'interprétation. En plus de la base de données ECM sur les polypes du côlon, utilisée dans les chapitres précédents, nous introduisons dans ce chapitre une nouvelle base de données ECM, sur l'œsophage de Barrett, pour montrer la généricité de notre méthode de reconnaissance. Des mosaïques ECM typiques de l'œsophage de Barrett sont présentées dans la figure 4.1. La performance de notre méthode de reconnaissance est évaluée indirectement en utilisant la classification binaire entre les cas pathologiques et non pathologiques. Comme nous l'avons montré dans le chapitre 2 sur la base de données des polypes du côlon, nous démontrons que sur la base de données de l'œsophage de Barrett notre méthode de reconnaissance surpasse plusieurs méthodes de l'état de l'art. A partir de nos résultats de reconnaissance, nous apprenons ensuite un prédicteur de* difficulté *en utilisant, comme vérité terrain sur la* difficulté *expérimentée, le pourcentage de faux diagnostics ECM parmi plusieurs médecins endoscopistes. Nos expériences démontrent que, pour les deux bases de données, il existe une corrélation significative entre la* difficulté *estimée à partir de la reconnaissance et la* difficulté *expérimentée par les endoscopistes.*

4.1 Introduction

Objective

Several training simulators have been proposed for the medical community, for example by Rhienmora et al. [Rhienmora 11] for dental surgery, by Pernod et

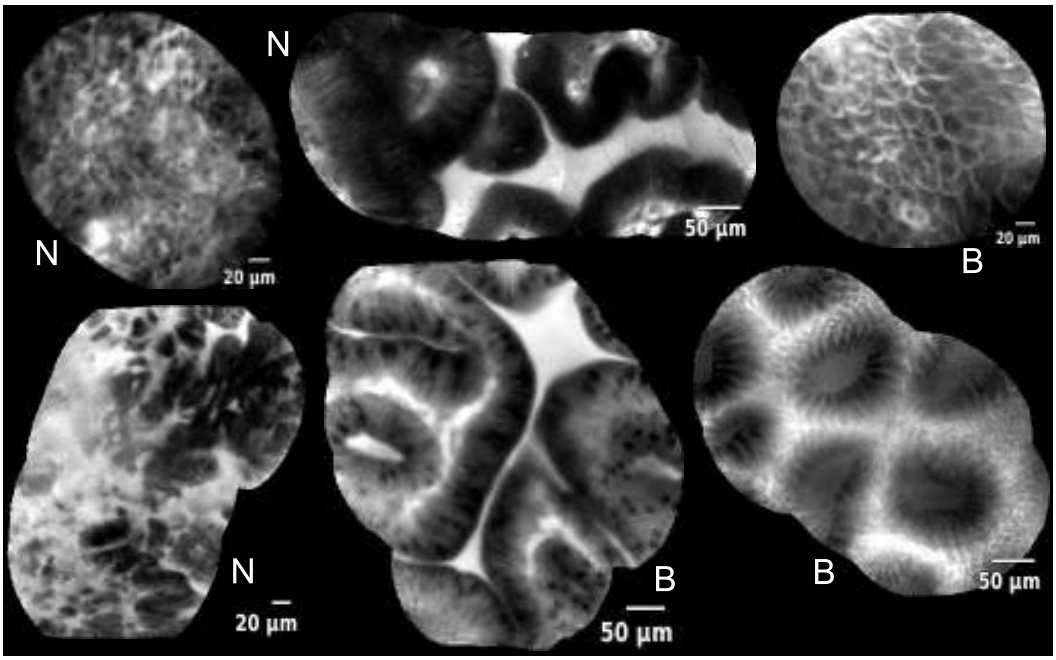


Figure 4.1: 6 **mosaic images of the Barrett database** (**B**: Benign, **N**: Neoplastic).

al. [Pernod 11] to support heart radio-frequency ablation, and by de Visser et al. [de Visser 10] for the next generation colonoscopy. Having a simulator for pCLE gesture training would be also interesting but gesture training is another problem, different from the one addressed in this chapter. Our focus is supporting the physician's interpretation.

The understanding of pathologies through the analysis of image sequences is a subjective learning experience which may be supported by modular training systems. Particularly, the early diagnosis of epithelial cancers from *in vivo* endomicroscopy is a challenging task for many non-expert endoscopists. Our objective is to develop a modular training system for endomicroscopy diagnosis, by adapting the *difficulty* level according to the expertise of the physician.

The training simulator, illustrated in Fig. 4.2, consists in a quiz. Given a level of *difficulty*, a pool of endomicroscopic videos whose average *difficulty* matches the current level is randomly chosen from the set of the training videos. By iterating this process with increasing levels of interpretation *difficulty*, the physician may be able to learn faster. The physician may also want to select the difficulty level in order to reinforce his/her diagnostic skills. For surgical skills, evidences of the efficiency of self-guided learning have been provided in the thesis of Brydges [Brydges 09], but further investigation is needed for the extension of learning effect analysis to diagnostic skills.

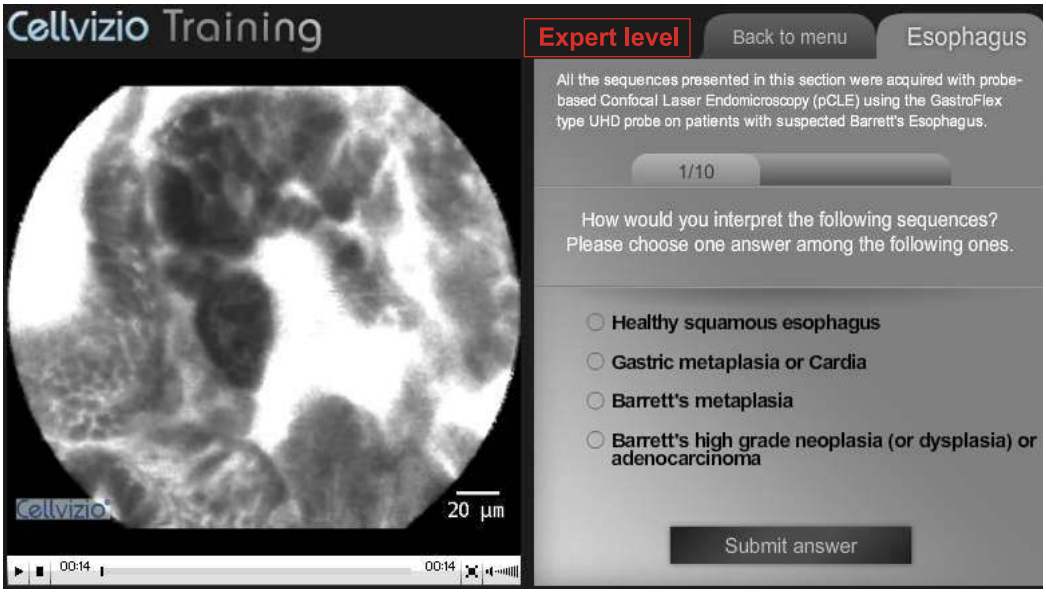


Figure 4.2: Screenshot of `www.cellvizio.net` Self-Learning Tool, with the added *difficulty* level information.

State of the art in estimating interpretation *difficulty*

Typical studies on query *difficulty* estimation consider textual queries, and not image queries. Besides, they usually do not predict the *difficulty* of the query interpretation but rather the *performance* of the query in order to estimate the quality of its retrieval results. However, given the tight analogy between text retrieval and image retrieval based on "visual words", the *difficulty* criteria used by these methods, most of which were presented in a survey by Hauff et al. [Hauff 08], may also be useful for our study. In particular, Zhao et al. [Zhao 08] estimated the performance of a textual query from similarity scores, but also from term frequency - inverse document frequency (TF-IDF) weights [Salton 88] extracted during the indexing time. In all these studies, the predictor validation process takes as ground truth an indicator of the performance of the retrieval system, such as the Average Precision (AP). Nevertheless, Scholer and Garcia [Scholer 09] demonstrated that the correlation between the estimated *difficulty* and the measured retrieval performance highly depends on the chosen retrieval system. Considering human performance in rating x-ray images as a ground truth, Schwaninger et al. [Schwaninger 07] proposed a statistical approach to estimate the image query *difficulty* solely from image measurements. Turpin and Scholer [Turpin 06] highlighted the fact that it is not easy to establish, for simple tasks like instance recall or question answering, a significant relationship between human performance and the performance of a retrieval system that uses precision-based measures to predict the query *difficulty*.

For our study, we consider videos as queries. We propose to learn a query *dif-*

ficulty predictor using relevant attributes from a Content-Based Video Retrieval (CBVR) method. We have two types of ground truth. For video retrieval, a diagnosis ground truth is the set of histological diagnoses of the biopsies associated to all the videos of the database. For interpretation *difficulty*, a *difficulty* ground truth is given by the percentage of false video-based diagnoses among several physicians on a subset of the video database. Histological diagnosis and video-based diagnosis both consist in differentiating benign from neoplastic (i.e. pathological) lesions. In these conditions, we aim at establishing a relationship between the physicians performance and our predictor.

Materials

Probe-based confocal laser endomicroscopy (pCLE) allows the endoscopist to image the epithelial tissue *in vivo*, at microscopic level with a miniprobe, and in real-time (18 frames per second) during an ongoing endoscopy.

The first pCLE database is of the **Colon** database used in 2, which contains 121 videos (36 benign, 85 neoplastic) split into 499 stable video sub-sequences (231 benign, 268 neoplastic). 11 endoscopists, among whose 3 experts and 8 non-experts, individually established a pCLE diagnosis on 63 videos (18 benign, 45 neoplastic) of the database. On the non-expert diagnosis database, interobserver agreement was assessed in the study of Buchner et al. [Buchner 09a], with an average accuracy of 72% (sensitivity 82%, specificity 53%). On the expert diagnosis database, Gomez et al. [Gomez 10] showed an interobserver agreement with an average accuracy of 75% (sensitivity 76%, specificity 72%). Thus, although pCLE is relatively new to many physicians, the learning curve pattern of pCLE in predicting neoplastic lesions was demonstrated with improved accuracies in time as observers' experience increased.

The second pCLE database is related to a different clinical application, namely the *Barrett's Esophagus*, and was provided by the multicentric "DONT BIOPCE" [DONT BIOPCE 10] study (Detection Of Neoplastic Tissue in Barrett's esophagus with In vivO Probe-based Confocal Endomicroscopy). Our resulting **Barrett** database includes 76 patients and contains 123 videos (62 benign, 61 neoplastic) split into 862 stable video sub-sequences (417 benign, 445 neoplastic). 21 endoscopists, among whose 9 experts and 12 non-experts, individually established a pCLE diagnosis on 20 videos (9 benign, 11 neoplastic) of the database.

For all these training videos, the pCLE diagnosis, either benign or neoplastic, is the same as the *gold standard* established by a pathologist after the histological review of biopsies acquired on the imaging spots.

4.2 Applying pCLE Retrieval to a New Database, the "Barrett's Esophagus"

In the current chapter, we apply for the first time our video retrieval method on a new pCLE database, different from the **Colon** database used in Chapter 2: the

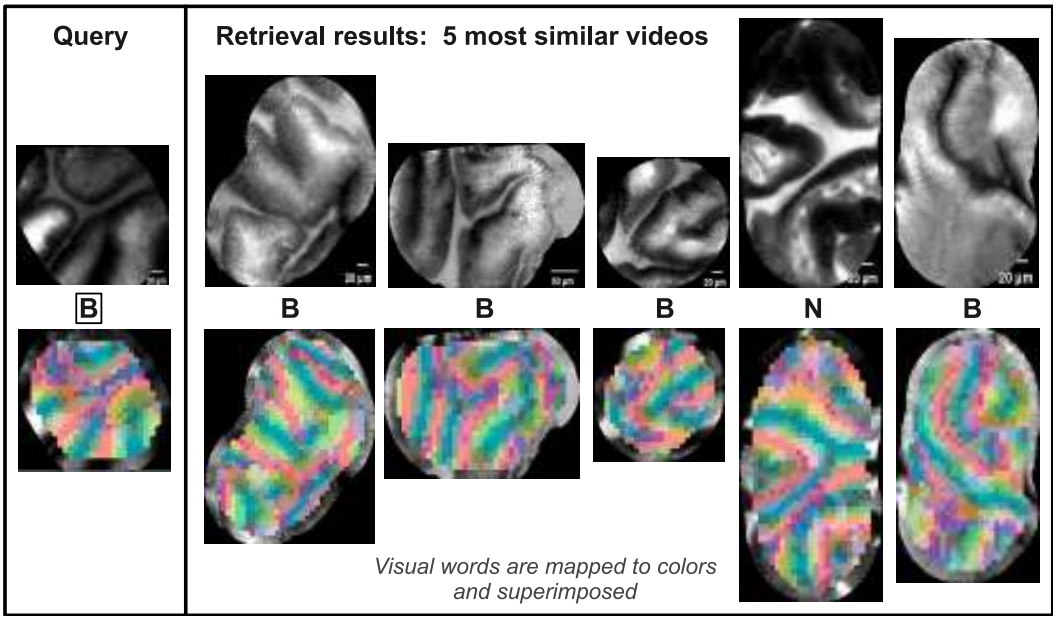


Figure 4.3: Typical video retrieval result of the "Dense-Scale-20" method applied to the Barrett's esophagus. B indicates Benign (i.e. non-neoplastic) and **N** Neoplastic (not present here). The pCLE videos are represented by their corresponding fused mosaic image built with non-rigid registration, and are shown together with their visual words. As observed on the **Colon** database in Chapter 2, the colored visual words are highlighting the geometrical structures in the mosaic images of the **Barrett** database.

Barrett database. We propose to use our dense video retrieval method presented in Chapter 2, which decomposes each video as a set of fused mosaic images built with non-rigid registration. Whereas we used disk regions of radius 60 pixels for the dense image description on the **Colon** database, we consider disk regions of radius 20 pixels in order to describe the images of the **Barrett** database whose discriminative patterns appear at a finer scale. As we did for the **Colon** database, we choose 20 pixels of grid spacing and $K = 100$ visual words for **Barrett** database, which will yield satisfying classification results given the relatively small size of the database. "Dense-Scale-60" is the retrieval method applied to the **Colon** database, and "Dense-Scale-20" the one applied to the **Barrett** database. As for the **Colon** database, the whole **Barrett** database is used both for training and testing, but leave-one-patient-out (LOPO) cross-validation is performed, as detailed in Chapter 2, for bias correction. A typical video retrieval result of the "Dense-Scale-20" method applied to the **Barrett** database can be qualitatively appreciated on Fig. 4.3.

For method comparison, we take as references the following CBIR methods, which we extended to CBVR by applying our signature summation technique: the HH-SIFT method presented by Zhang et al. [Zhang 07] a sparse detector, the stan-

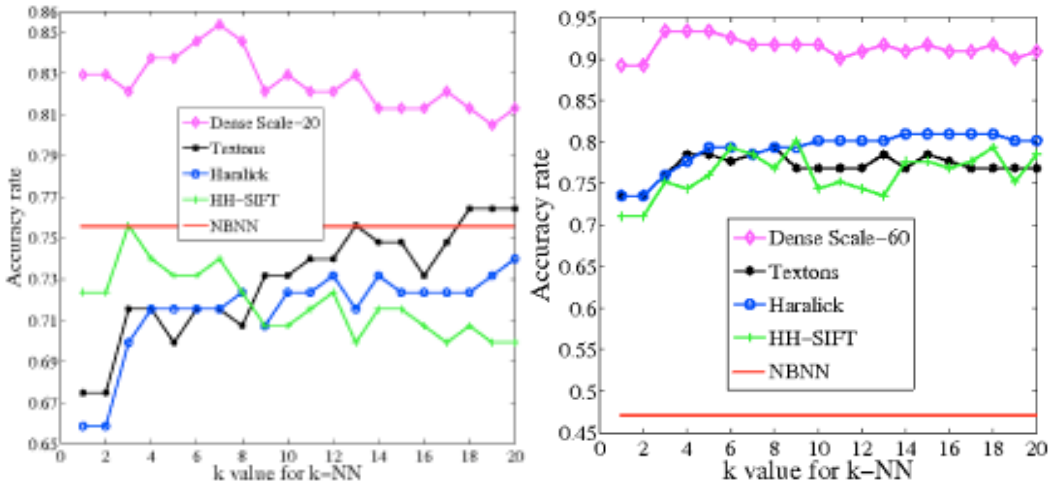


Figure 4.4: Left: Method comparison for the LOPO classification of pCLE videos on the **Barrett** database, with the default values $\theta = 0$ and $\theta_{\text{NBNN}} = 1$. **Right:** Method comparison for the LOPO classification of pCLE videos on the **Colon** database, with the default values $\theta = 0$ and $\theta_{\text{NBNN}} = 1$.

dard approach of Haralick features, the texture retrieval Textons method of Leung and Malik [Leung 01], and an efficient image classification method presented by Boiman et al. [Boiman 08], referred as NBNN, that uses no clustering. As an indirect means to evaluate retrieval performance we use k-nearest neighbor classification, for which we consider two pathological classes, benign (vote $= -1$) and neoplastic (vote $= +1$).

The accuracy results of video classification on the **Barrett** database are presented in Fig. 4.4. In agreement with the ROC curves shown in Fig. 4.5, the accuracy results obtained on the **Colon** database are even better. Our retrieval method outperforms all the compared methods with a gain of accuracy greater than 12 percentage points on the **Colon** database, and greater than 9 percentage points on the **Barrett** database. McNemar's tests show that, when the number k of neighbors is fixed, the improvement of our method with respect to all others is statistically significant: p-value < 0.011 for $k \in [1, 10]$ on the **Colon** database and p-value < 0.043 for $k \in [1, 2] \cup [4, 8]$ on the **Barrett** database. This shows the genericity of our retrieval method, which is successfully applied to two different clinical application, with: 93.4% of accuracy (sensitivity 95.3%, specificity 88.9%) at $k = 3$ neighbors on the **Colon** database, and 85.4% of accuracy (sensitivity 90.2%, specificity 80.7%) at $k = 7$ neighbors on the **Barrett** database.

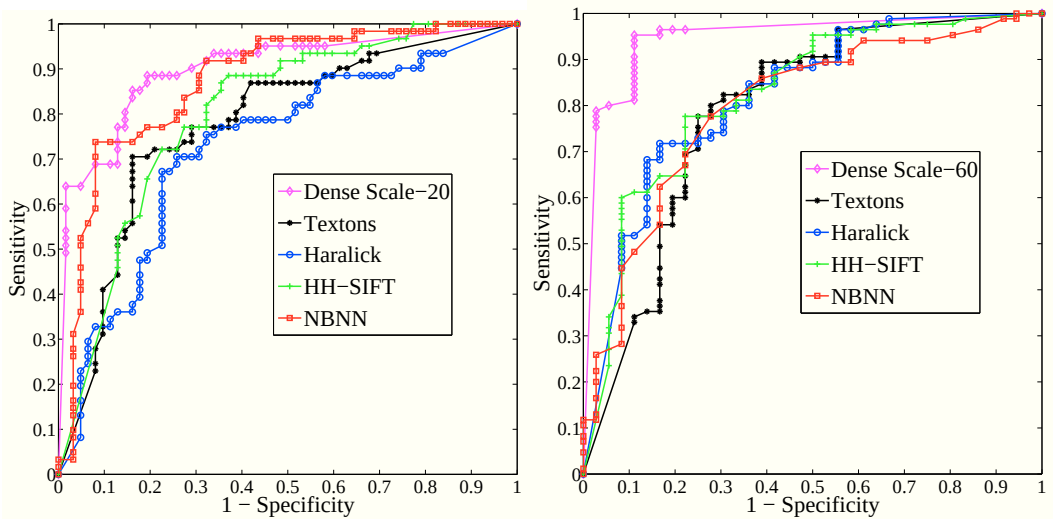


Figure 4.5: Left: ROC curves at $k = 5$ neighbors from LOPO video classification on the **Barrett** database, with $\theta \in [-1, 1]$ and $\theta_{\text{NBNN}} \in [0, +\infty[$. **Right:** ROC curves at $k = 5$ neighbors from LOPO video classification on the **Colon** database, with $\theta \in [-1, 1]$ and $\theta_{\text{NBNN}} \in [0, +\infty[$. θ and θ_{NBNN}, introduced in Chapter 2, trade off the cost of false positives and false negatives.

4.3 Estimating the Interpretation *Difficulty*

For *difficulty* estimation, our ground truth is given by the percentage, for each query video, of false diagnoses among the physicians. As the understanding of video diagnosis by the physicians is driven by similarity-based reasoning, it makes sense to predict the query *difficulty* based on similarity results of video retrieval.

To learn a *difficulty* predictor, our idea is to exploit, as relevant attributes, the results of our video retrieval method applied to the training database. Potential relevant attributes are the class $C_q \in \{-1, +1\}$ of the video query q, the classes $C^{j \in \{1,..,k\}} \in \{-1, +1\}$ of its k nearest neighbors and the similarity distances from them to the query. Given the small number of videos tested by the involved physicians, too many attributes for *difficulty* learning may lead to over-fitting. For this reason, we decided to extract, from the retrieval results, one efficient and intuitive *difficulty* attribute α which reflects the contextual discrepancies between the video query and its similarity neighborhood. For each query video, we thus considered the retrieval error between the class of the query and a weighted average of its neighbors' votes:

$$\alpha_q = 1 - C_q \frac{\sum_{j=1}^{k} C^j \; z_{C^j}}{\sum_{j=1}^{k} z_{C^j}} \tag{4.1}$$

where $z_{-1} = 1$ and z_{+1} is a constant weight applied to the neoplastic votes. By default $z_{+1} = 1$, which corresponds to putting the same weight to neoplastic and non-neoplastic votes. Introducing z_{+1} allows us to take into account the possible

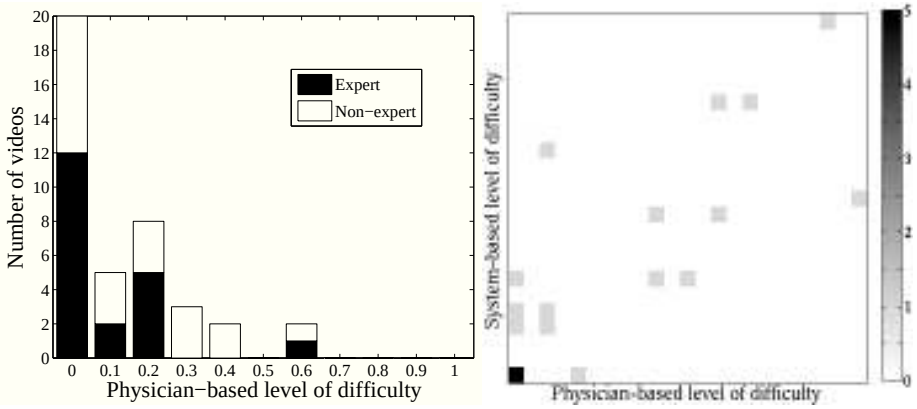


Figure 4.6: Left: Difficulty ground-truth histograms on the **Barrett** database. **Right:** Joint histograms on the **Barrett** database; x-axis is the *difficulty* experienced by all the physicians and y-axis is our estimated *difficulty*. On the **Barrett** database, 21 physicians, 9 expert and 12 non expert, individually diagnosed 20 videos.

emphasis of neoplastic votes with respect to the benign votes. Our query *difficulty* predictor P is thus defined as $P(q) = \alpha_q$ for each query video q. Its relevance can be evaluated by a simple correlation measure between the estimated difficulties of all tested videos and their ground-truth values. In this case, as there is no learning process, cross-validation is not necessary.

4.4 Results of the *Difficulty* Estimation Method

4.4.1 Results on the Barrett database

We experiment our *difficulty* predictor presented in Section 4.3 on the **Barrett** database. The best Pearson correlation coefficients are obtained with $k = 10$ neighbors and a neoplastic weight $z_{+1} = 0.4$. The correlation coefficients reach 0.78 with respect to the *difficulty* experienced by all the physicians, 0.63 (resp. 0.80) with respect to the *difficulty* experienced by the experts only (resp. the non-experts only). The corresponding joint histogram is presented in Fig. 4.6, along with the histogram of the *difficulty* ground-truth values.

Permutation tests demonstrate that there a significant correlation (p-value $<$ 0.005) between the ground truth and our proposed *difficulty* estimation, which confirms the efficiency of our retrieval-based attribute for intuitive *difficulty* estimation. We refer the reader to the Appendix A for a detailed description of the permutation test.

Because the video subset for which we have the *difficulty* ground truth is limited to 20 videos, which is insufficient for learning purpose, we cannot perform the learning of the interpretation *difficulty* on the **Barrett** database. However, *difficulty* learning will be explored on the **Colon** database, for which the ground-truth data contains three times more videos.

4.4.2 Results on the Colon database

On the **Colon** database, the *difficulty* estimation results are not as good as on the **Barrett** database. With $k = 10$ neighbors and a neoplastic weight $z_{+1} = 6$, the Pearson correlation coefficients reach 0.45 with respect to the *difficulty* experienced by all the physicians, 0.30 (resp. 0.45) whit respect to the *difficulty* experienced by the experts only (resp. the non-experts only). Permutation tests demonstrate, however, that there a significant correlation (p-value < 0.005) between this estimated *difficulty* and the experienced difficulty.

In order to improve the correlation results, we propose to investigate a machine learning-based approach, which will need more relevant attributes. As the video subset for which we have the *difficulty* ground truth is relatively small, i.e. 63 videos, we decide to add only one attribute β that measures the intrinsic ambiguity of the video query with respect to the two pathological classes. We then learn the *difficulty* predictor from the two attributes α and β by using a robust linear regression model. Our intrinsic attribute β reflects the standard deviation of the "signed" discriminative power of the query signature, with respect to the benign and the neoplastic classes:

$$\beta = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(w_i \; g_s(i) - \sum_{j=1}^{K} w_j \; g_s(j))^2} \tag{4.2}$$

where w_i is the frequency of the i^{th} visual word in the video query and $g_s(i)$ is its "signed" discriminative power given by the adapted Fisher criterion:

$$g_s(i) = \frac{1}{2} \; \frac{(\mu_{-1}(i) - \mu_{+1}(i)) \; |\mu_{-1}(i) - \mu_{+1}(i)|}{\sigma_{-1}(i)^2 + \sigma_{+1}(i)^2} \tag{4.3}$$

where $\mu_C(i)$ and $\sigma_C(i)$ are respectively the mean and the variance of the frequency distribution of the i^{th} visual word in the videos belonging to class C, with $C \in \{-1, +1\}$.

The Pearson correlation coefficients obtained by the robust linear regression model with cross-validation reach 0.48 when learning from the *difficulty* experienced by all the physicians, 0.33 (resp. 0.47) when learning from the *difficulty* experienced by the experts only (resp. the non-experts only). Even if these correlation results are less convincing than those obtained on the **Barrett** database, the correlation coefficient has been improved by learning. The corresponding joint histogram is presented in Fig. 4.7. According to the permutation tests, the correlation between this learned *difficulty* and the experienced *difficulty* remains statistically significant (p-value of < 0.005).

To automate the optimal attributes selection and to explore potentially more relevant attributes for *difficulty* estimation, further experiments based on model selection need to be investigated, for example using the Akaike information criterion [Akaike 74] or the Bayesian information criterion [Schwarz 78]. Besides, selection criteria commonly used in active learning [Hoi 08] may help to provide a better *difficulty* estimation. However, this requires larger training databases.

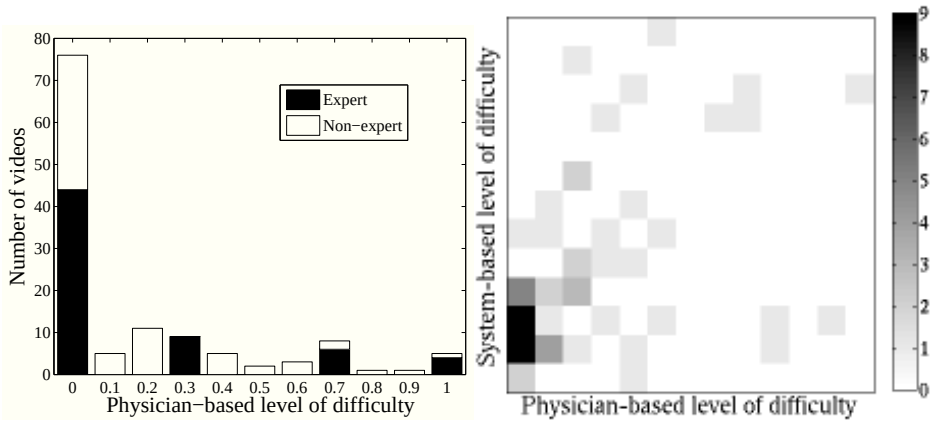


Figure 4.7: Left: Difficulty ground-truth histograms on the **Colon** database. **Right:** Joint histograms on the **Colon** database; x-axis is the *difficulty* experienced by all the physicians and y-axis is our estimated *difficulty*. On the **Colon** database, 11 physicians, 3 expert and 8 non expert, individually diagnosed 63 videos.

4.5 Conclusion

To our knowledge this study proposes the first approach to learn, for endomicroscopy training, the interpretation *difficulty* experienced by human experts, based on an original method of Content-Based Video Retrieval. Our experiments have demonstrated that there is a significant relationship between our retrieval-based *difficulty* estimation and the *difficulty* experienced by the physicians. Moreover, we showed the promising genericity of our *difficulty* estimation method by applying it on two different clinical databases, one on the Barrett's Esophagus and the other on colonic polyps. Our method could also be potentially applied to other imaging applications.

On one hand we have the diagnosis ground truth for all the videos belonging to our two large databases, on the other hand we have the *difficulty* ground truth on a small subset of each database. The method proposed in this work can then be used to estimate the interpretation *difficulty* on the remaining videos. It is worth noticing that, if no *difficulty* ground truth is available, or if it is not large enough for learning, as it is the case for the database on Barrett's esophagus, we are still able to estimate the interpretation *difficulty* of any video. The full pCLE databases, completed with the *difficulty* estimation, could then be used in a self-training simulator that features *difficulty* level selection. For example, given a *difficulty* level x, query videos can be randomly drawn by the simulator according to a Gaussian probability distribution centered at x and of suitable variance. Such a structured training simulator should make endomicroscopy training more relevant. Finally, a clinical validation would be required to see whether the self-training simulator could help shorten the physician learning curve.

Chapter 5

Learning Semantic and Visual Similarity between Endomicroscopy Videos

Table of Contents

Based on: [André 11c] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *Learning semantic and visual similarity for endomicroscopy video retrieval*. 2011. Article in submission. **Additional material available in** [André 11d].

Traditional CBIR systems only deliver visual outputs, i.e. images having a similar appearance to the query, which is not directly interpretable by the physicians. Our objective is to provide a system for endomicroscopy video retrieval which delivers both visual and semantic outputs that are consistent with each other. In Chapter 2, we developed the "Dense-Sift" method for endomicroscopy retrieval that computes, for each video represented as a set of fused mosaic images, a single visual *signature. In this study, we first leverage semantic ground-truth data to transform*

these visual *signatures into* semantic *signatures that reflect how much the presence of each semantic concept is expressed by the visual words describing the videos. Using cross-validation, we demonstrate that our visual-word-based* semantic *signatures enable a recall performance which is slightly lower than that of the* visual *signatures computed by "Dense-Sift". Nevertheless, the relevance of the* semantic *signatures is shown by the fact that their recall performance remains significantly higher than those of several state-of-the-art methods in CBIR. In a second step, we propose to improve retrieval relevance by learning, from a perceived similarity ground truth, an adjusted similarity distance. Our distance learning method allows to improve, with statistical significance, the correlation with the perceived similarity. Although* semantic *signatures and* visual *signatures have comparable performances in terms of correlation with the perceived similarity, the* semantic *signatures communicate high-level medical knowledge while being consistent with the low-level* visual *signatures and much shorter than them. Our resulting retrieval system is efficient in providing both visual and semantic information that are correlated with each other and clinically interpretable by the endoscopists.*

French summary

Les systèmes de CBIR traditionnels ne délivrent que des sorties visuelles, c'est à dire des images ayant une apparence similaire à la requête. Or une sortie visuelle n'est pas directement interprétable par le médecin dans son propre langage. Notre objectif est de fournir un système de reconnaissance de vidéos endomicroscopiques délivrant des sorties à la fois visuelles et sémantiques qui sont consistantes entre elles. Dans le chapitre 2, nous avons développé la méthode "Dense-Sift" pour la reconnaissance en endomicroscopie. Celle-ci calcule, pour chaque vidéo représentée par un ensemble de mosaïques, une unique signature visuelle. *Dans cette étude, nous exploitons tout d'abord une vérité terrain sémantique pour transformer ces signatures* visuelles *en signatures* sémantiques *qui reflètent à quel point la présence de chaque concept sémantique est exprimée par les mots visuels décrivant les vidéos. En utilisant la validation croisée, nous démontrons que nos signatures* sémantiques *fondées sur les mots visuels permettent d'obtenir une performance de rappel légèrement inférieure à celle des signatures* visuelles *calculées par la méthode "Dense-Sift", mais supérieure de manière significative aux performances de rappel obtenues par plusieurs méthodes de l'état de l'art en CBIR. Dans un deuxième temps, nous proposons d'améliorer la pertinence de la reconnaissance an apprenant, à partir d'une vérité terrain sur la similarité perçue, une distance de similarité ajustée. Notre méthode d'apprentissage de la distance permet d'améliorer, de manière significative, la corrélation avec la similarité perçue. Bien que les signatures* sémantiques *et les signatures* visuelles *aient des performances comparables en termes de corrélation avec la similarité perçue, les signatures* sémantiques *communiquent des connaissances médicales de haut niveau, tout en étant consistantes avec les signatures*

visuelles *de bas niveau et beaucoup plus courtes qu'elles. Notre système de reconnaissance final est efficace dans l'extraction d'informations visuelles et sémantiques corrélées entre elles et interprétables sur le plan clinique par les endoscopistes.*

5.1 Introduction

The expanding application of Content-Based Image Retrieval (CBIR) methods of computer vision in the medical diagnosis field faces the semantic gap, which was pointed out by Smeulders et al. in [Smeulders 00] and by Akgül et al. in [Akgül 11], as a critical issue. In CBIR, the semantic gap is the disconnection between the reproducible computational representation of low-level visual features in images and the context-dependent formulation of high-level knowledge, or semantics, to interpret these images. Two medical images being highly similar in appearance may have contradictory semantic annotations. So a CBIR system, which would be only based on visual content, might lead the physician toward a false diagnosis. Conversely, two medical images having exactly the same semantic annotations may look visually dissimilar. So a CBIR system, for which the semantics of the query is unknown, might not retrieve all clinically relevant images. In fact, when interpreting a new image for diagnostic purposes, the physician uses similarity-based reasoning, where *similarity* includes both visual features and semantic concepts. To mimic this process, we aim at capturing the visual content of images using the Bag-of-Visual-Words (BoW) method, and at estimating the expressive power of visual words with respect to multiple semantic concepts. The consistency of the induced visual-word-based *semantic* retrieval could then be tested against perceived similarity ground truth.

Our medical application is the retrieval of probe-based Confocal Laser Endomicroscopy (pCLE) videos to support the early diagnosis of colonic cancers. pCLE is a recent imaging technology that enables the endoscopist to acquire *in vivo* microscopic video sequences of the epithelium, and thus to establish a diagnosis in real time. In particular, the *in vivo* diagnosis of colonic polyps using pCLE is still challenging for many endoscopists, because of the high variability in the appearance of pCLE videos and the presence of atypical cases such as serrated adenoma [Khalid 09]. Fig. 5.1 shows an illustration of the semantic gap in endomicroscopy retrieval, with several examples of mosaic images extracted from pCLE videos. In Chapter 2 we have developed a dense BoW method, called "Dense-Sift", for the content-based retrieval of pCLE videos. We showed that, when evaluated in terms of pathological classification of pCLE videos, "Dense-Sift" significantly outperforms several state-of-the-art CBIR methods. Parts of this chapter are extensions of a preliminary study [André 11d] where we explored pCLE retrieval evaluation and distance learning in terms of perceived visual similarity. Here, our objective is to learn the pCLE similarity distance both in terms of visual appearance and semantic annotations, in order to provide the endoscopists with semantic insight into the retrieval results.

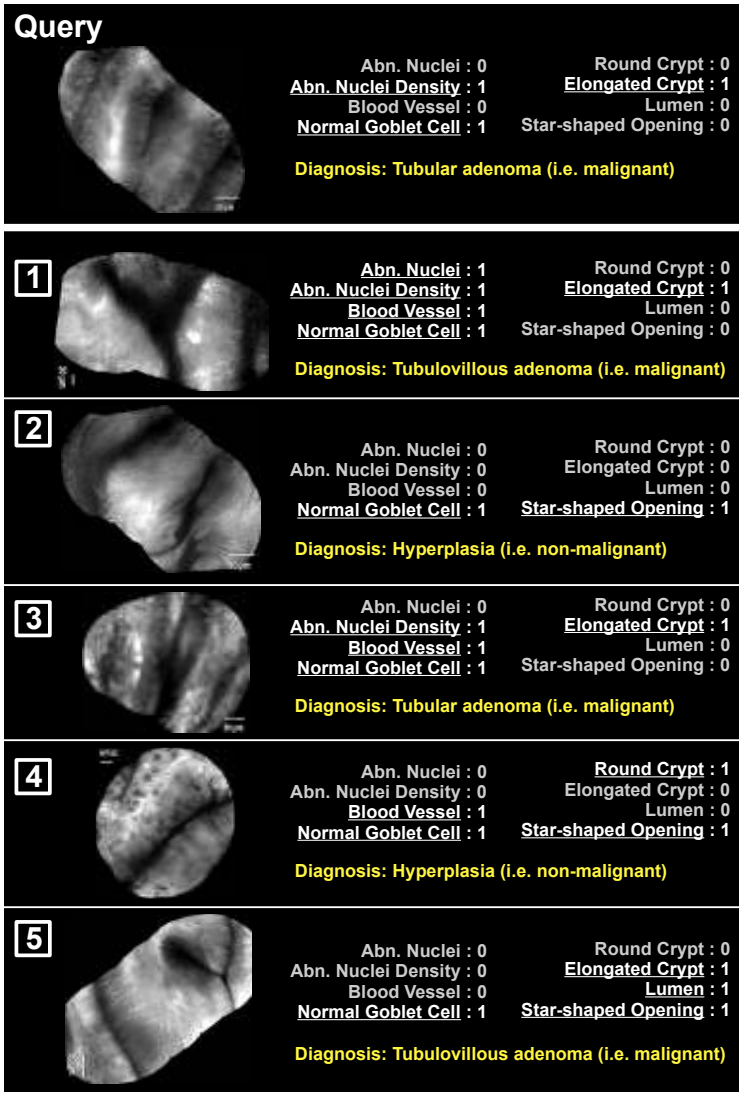


Figure 5.1: Illustration of the semantic gap: content-based retrieval of visually similar pCLE videos having dissimilar semantic annotations. The 5 most similar pCLE videos are retrieved by the "Dense-Sift" method that only relies on visual features. Semantic concepts which were annotated as present in a given video are underlined. For each video, the pathological diagnosis, either malignant or non-malignant, is indicated below the semantic concepts. For illustration purposes, videos are represented by mosaic images.

To this purpose, we introduce in Section 5.2 two new types of ground truth, which are different from the diagnosis ground truth and the *difficulty* ground truth used in the previous chapters. The first new type of ground truth contains visual similarities perceived by endoscopists between pCLE videos, evaluated on a four-point Likert scale. The second new type of ground truth contains multiple binary semantic concepts identified by experts in pCLE videos. These eight binary

concepts, illustrated in Fig. 5.2, have been defined to support the *in vivo* pCLE diagnosis of colonic polyps. From the *visual* signatures computed by our "Dense-Sift" retrieval method, and from the semantic ground truth, we build visual-word-based *semantic* signatures using a Fisher-based approach detailed in Section 5.4. We evaluate the relevance of the resulting *semantic* signatures, first from the sole semantic point of view, with ROC curves showing classification performances for each semantic concept, and then from the perceptual point of view, with *sparse recall* curves showing the ability of the induced retrieval system to capture video pairs perceived as *very similar*. Retrieval performance is also evaluated by measuring the correlation of the induced similarity distance with the perceived similarity ground truth. In order to improve retrieval relevance, we propose in Section 5.5 a method to learn an adjusted similarity distance from the perceived similarity ground truth. A linear transformation of video signatures is optimized, that minimizes a margin-based cost function differentiating *very similar* video pairs from the others. The results shown in Section 5.6 show that the visual-word-based *semantic* signatures yield a recall performance which is slightly lower than that of the original *visual* signatures computed by "Dense-Sift", but significantly higher than those of several state-of-the-art methods in CBIR. In terms of correlation with the perceived similarity, the retrieval performance of *semantic* signatures is better, with statistical significance, than those of the state-of-the-art methods, and comparable to that of the original *visual* signatures. For both *semantic* signatures and *visual* signatures, the distance learning method allows to improve, with statistical significance, the correlation with the perceived similarity. Our resulting pCLE retrieval system, of which visual and semantic outputs are consistent with each other, should better assist the endoscopist in establishing a pCLE diagnosis.

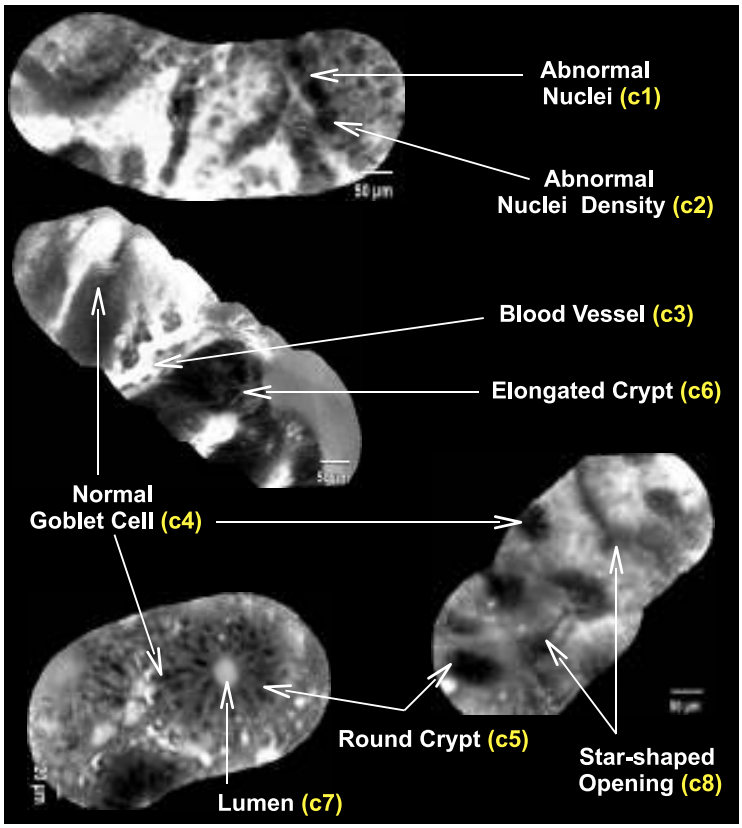


Figure 5.2: Examples of training pCLE videos represented by mosaic images and annotated with the 8 semantic concepts. The two mosaics on the top show neoplastic (i.e. malignant) colonic polyps, while the two mosaics on the bottom show non-neoplastic (i.e. non-malignant) colonic polyps.

5.2 Ground Truth for Perceived Visual Similarity and for Semantics

5.2.1 pCLE database

Our video database is a large subset of the *Colonic Polyp* database used in Chapter 2, from which all the polyps having incomplete semantic data have been excluded. The resulting database contains 118 pCLE videos of colonic polyps that were acquired from 66 patients. The lengths of the acquired pCLE videos range from 1 second to 4 minutes. Each pCLE video is represented as a set of fused mosaic images built with the video-mosaicing technique of Vercauteren et al. [Vercauteren 06]. Dabizzi et al. [Dabizzi 11] and De Palma et al. [De Palma 10] recently showed that pCLE mosaics have the potential to replace pCLE videos for a comparable diagnosis accuracy and a significantly shorter interpretation time. For this reason, pCLE mosaic images will not only be used as input for our retrieval system, but also as retrieval outputs attached to the extracted similar videos.

5.2.2 Ground Truth for Perceived Visual Similarity

To generate a pairwise similarity ground truth between pCLE videos, we designed an online survey tool, called VSS [VSS 11], which is available online at `http://smartatlas.maunakeatech.com` for testing purpose (login: MICCAI-User, password: MICCAI2011). The VSS tool allows multiple observers, who are fully blinded to the video metadata such as the pCLE diagnosis, to qualitatively estimate the perceived visual similarity degree between videos. For each video couple, the following four-point Likert scale are proposed by the survey tool: *very dissimilar*, *rather dissimilar*, *rather similar* and *very similar*. Because interpreting whole video sequences is time consuming, the VSS supports this task by making available both the whole video content and for each video, the corresponding set of static mosaic images providing a visual summary. A screenshot of the online VSS tool is shown in Fig. 5.3. Each scoring process, as illustrated in Fig. 5.4, is characterized by the random drawing of 3 video couples (I_0, I_1), (I_0, I_2) and (I_0, I_3), where the candidate videos I_1, I_2 and I_3 belong to patients that are different from the patient of the reference video I_0, in order to exclude any patient-related biases. 17 observers, ranging from middle expert to expert in pCLE diagnosis, performed as many scoring processes as they could. Our generated ground truth can be represented as an graph where the nodes are the videos and where each couple of videos may be connected by zero, one or several edges representing the similarity scores. As less than 1% of these video couples were scored by more than 4 distinct observers, it was not relevant to measure inter-observer variability. In total, $4,836$ similarity scores were given for $2,178$ distinct video couples. Thus 16.2% of all $13,434$ distinct video couples were scored. Compared to our preliminary study [André 11d] where 14.5% of all possible video couples were scored, the perceived similarity ground truth was enriched for this study in order to better differentiate potentially *very similar* video pairs from the others, a goal which is closer to our retrieval purpose.

If the video couples were randomly drawn with a uniform non-informative prior by the VSS tool, we would have drawn much more video pairs perceived as *dissimilar* than video pairs perceived as *very similar*. The resulting perceived similarity ground truth would have been too far from our clinical application which aims at extracting highly similar videos. For this reason, we use the *a priori* similarity distance d_{Vis} computed by the "Dense-Sift" method to enable two modes for the drawing of video pairs: while the first mode biases the drawing by the use of the similarity distance d_{prior} computed by "Dense-Sift", the second mode only allows the drawing of nearest neighbors defined by "Dense-Sift".

- In the first mode, the probability of drawing a video couple (I_i, I_j) is proportional to the inverse of the density of $d_{prior}(I_i, I_j)$.
- In the second mode, the video I_j is one of the 5 nearest neighbors of the video I_i according to the retrieval distance d_{Vis}.

A total of $3,801$ similarity scores was recorded with the first mode, and $1,035$ with the second mode.

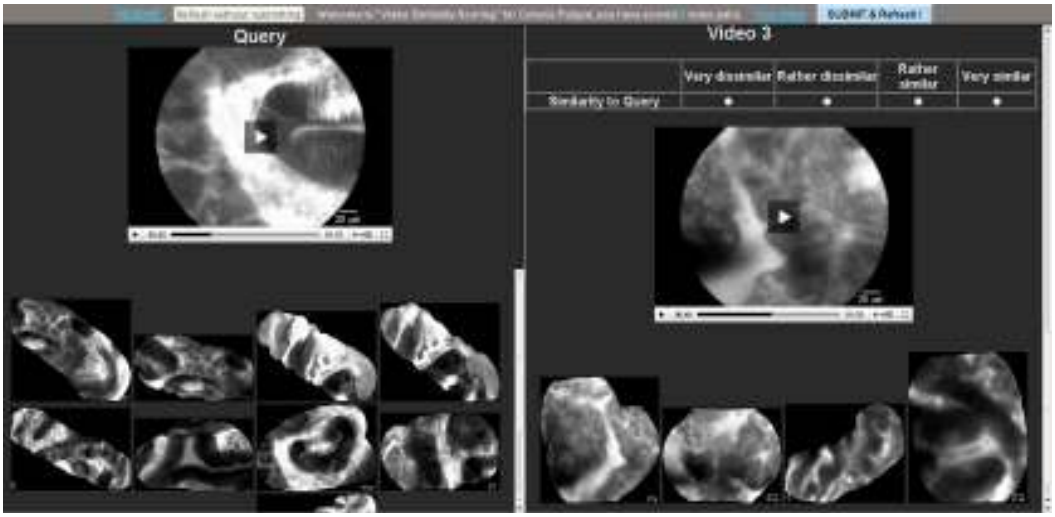


Figure 5.3: Screenshot of the online VSS tool: visual similarity scoring between the reference Video Query and the candidate Video 3. Each video is summarized by a set of mosaic images. Vertical scrollbars allow to see more mosaics of the Video Query (on the left), or the two other candidate videos 1 and 2 with their mosaics (on the right).

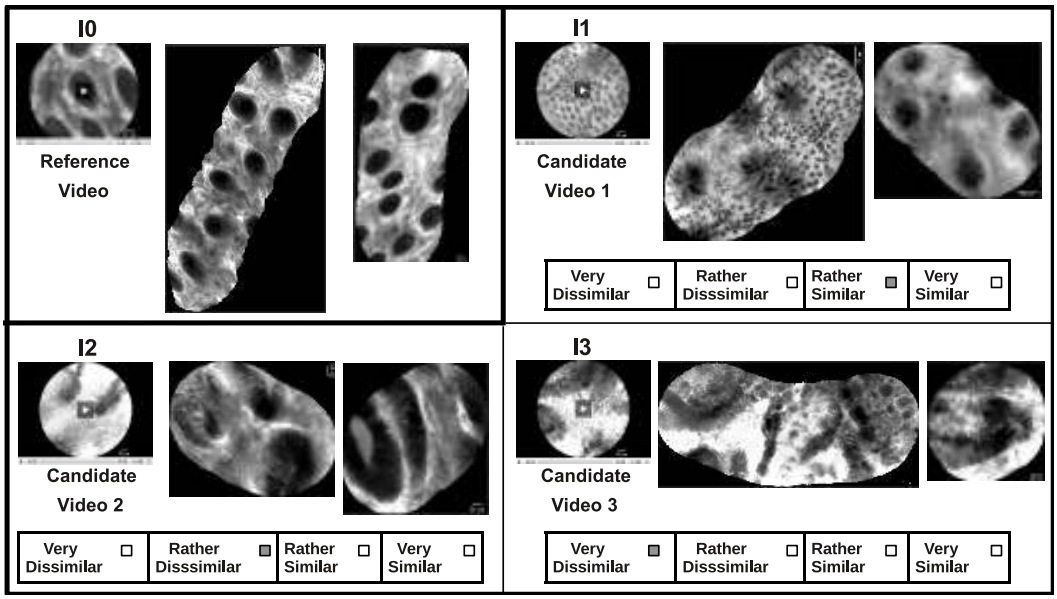


Figure 5.4: Schematic outline of the online "Visual Similarity Scoring" tool showing the example of a scoring process, where 3 video couples (I_0, I_1), (I_0, I_2) and (I_0, I_3) are proposed. Each video is summarized by a set of mosaic images.

Semantic concept	Indicator of representativity
c1. abnormal nuclei	46.6 %
c2. abnormal nuclei density	63.6 %
c3. blood vessel	47.5 %
c4. normal goblet cell	72.0 %
c5. round crypt	47.5 %
c6. elongated crypt	64.4 %
c7. lumen	27.1 %
c8. star-shaped opening	18.6 %

Table 5.1: Indicator of the representativity of each semantic concept. The representativity of each semantic concept is measured by the percentage of the videos in the database where the concept is annotated as visible.

Although the resulting similarity graph remains very sparse, we will show in Section 5.6 that it constitutes a valuable ground-truth database for retrieval evaluation and for perceived similarity learning.

5.2.3 Ground Truth for Semantic Concepts

All the acquired pCLE videos were manually annotated with $M = 8$ binary semantic concepts describing the observed colonic polyps. These concepts are illustrated on pCLE mosaic images in Fig. 5.2. In a given pCLE video, each semantic concept is defined as either visible, potentially several times, or not visible at all in the video. The first two concepts, *abnormal nuclei* (c_1) and *abnormal nuclei density* (c_2), which are the most difficult to identify, were annotated by two expert endoscopists. With the support of the modified Mainz criteria identified by Kiesslich et al. [Kiesslich 04] six other concepts were annotated: *blood vessel* (c_3), *normal goblet cell* (c_4), *round crypt* (c_5), *elongated crypt* (c_6), *lumen* (c_7) and *star-shaped opening* (c_8). If the semantic j^{th} concept is visible in the video then $c_j = 1$ else $c_j = 0$. Table 5.1 shows, for each semantic concept, the percentage of the videos in the database where the concept is annotated as visible.

5.3 From pCLE Videos to Visual Words

"Dense-Sift" is the pCLE video retrieval method developed in Chapter 2 that uses a dense description based on a disk radius of 60 pixels and that decomposes a video as a set of fused mosaic images. As a result, "Dense-Sift" computes a visual word signature $\mathscr{S}_{\text{Vis}}(I) = (w_1^I, ..., w_K^I)$ for each pCLE video I, where w_i^I is the frequency of the i^{th} visual word in the video I. We define the visual similarity distance $d_{\text{Vis}}(I, J)$ between two videos I and J as the χ^2 pseudo-distance between

their visual word signatures computed by "Dense-Sift":

$$\begin{aligned} d_{\text{Vis}}(I,J) &= \chi^2(\mathscr{S}_{\text{Vis}}(I), \mathscr{S}_{\text{Vis}}(J)) \\ &= \frac{1}{2} \sum_{i\in\{1,..,K\}, w_i^I w_i^J > 0} \frac{(w_i^I - w_i^J)^2}{w_i^I + w_i^J} \end{aligned} \tag{5.1}$$

As in Chapter 2, we will compare the retrieval performances of our "Dense-Sift" method with the following three competitive CBIR methods: "HH-Sift" of Zhang et al. [Zhang 07], "Textons" of Leung and Malik [Leung 01], and "Haralick" [Haralick 79]. Our "Dense-Sift" method was proved in Chapter 2 to be the best method in terms of pathological classification of pCLE videos. "Dense-Sift" will also be proved to be the best method in terms of correlation with the perceived visual similarity, as shown in Section 5.6. For these reasons, we decide to build the *semantic* signatures of pCLE videos from the *visual* signatures computed by "Dense-Sift".

5.4 From Visual Words to Semantic Signatures

Among the approaches in bridging the semantic gap, recent methods based on random-walk processes on visual-semantic graphs were proposed by Poblete et al. [Poblete 10] and by Ma et al. [Ma 10]. Latent semantic indexing approaches have also been investigated, for example by Caicedo et al. [Caicedo 10] to improve medical image retrieval. Rasiwasia et al. [Rasiwasia 07, Rasiwasia 10] proposed a probabilistic method which we consider as a reference method for performing a semantic retrieval which is based on visual features. In particular, their approach estimates for each semantic concept the probability that, given a visual feature vector in an image, the semantic concept is present in the image. In [Kwitt 11], Kwitt et al. recently applied this method for learning pit pattern concepts in endoscopic images of colonic polyps. These pit pattern concepts at the macroscopic level can be seen as corresponding to our semantic concepts at the microscopic level. In order to learn semantic concepts from visual words in endomicroscopic videos, we propose a rather simple method providing satisfactory results. The application of a probabilistic method such as the one in [Rasiwasia 07] on our data was not successful, certainly because of our relatively small sample size, but we plan to further investigate it. Our proposed method is a Fisher-based approach that estimates the expressive power of each of the K visual words with respect to each of the M semantic concepts.

Let D^{train} be the set of training videos. Given the i^{th} visual word and the j^{th} semantic concept, we estimate the discriminative power of the i^{th} visual word with respect to j^{th} semantic concept using the *signed* Fisher criterion:

$$F_{i,j} = \frac{\mu_1(i,j) - \mu_0(i,j)}{\sigma_1^2(i,j) + \sigma_0^2(i,j)} \tag{5.2}$$

where $\mu_p(i,j)$ (resp. $\sigma_p^2(i,j)$) is the mean (resp. the variance) of the visual word frequencies $\{w_i^I,\ c_j^I = p,\ I \in D^{train}\}$ with $p = 0$ or $p = 1$. We call F the re-

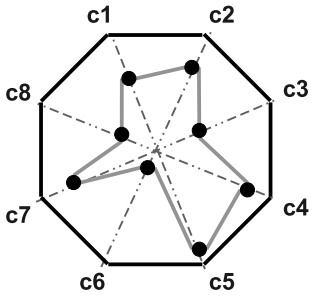


Figure 5.5: An example of a star plot based on the 8 semantic concepts. The coordinate value along the j^{th} radius corresponds to the normalized value of the *semantic* signature at the j^{th} concept.

sulting matrix of Fisher's weights. Given a video I of *visual* signature $\mathscr{S}_{\text{Vis}}(I) = (w_1^I, ..., w_K^I)$, we define the *semantic weight* of I with respect to j^{th} semantic concept as the following linear combination: $s_j^I = \sum_{i=1}^{K} F_{i,j} w_i^I$. Thus, the transformation from the *visual* signature $\mathscr{S}_{\text{Vis}}(I)$ into its visual-word-based *semantic* signature $\mathscr{S}_{\text{Sem}}(I) = (s_1^I, ..., s_M^I)$ is given by the equation:

$$\mathscr{S}_{\text{Sem}}(I) = F^T \mathscr{S}_{\text{Vis}}(I) \tag{5.3}$$

The signed value s_j^I reflects how much the presence of the j^{th} semantic concept is expressed by the visual words describing the video I. Finally, a visual-word-based *semantic* similarity distance between two videos I and J can be defined for example using the L^2 norm:

$$d_{\text{Sem}}(I, J) = ||\mathscr{S}_{\text{Sem}}(I) - \mathscr{S}_{\text{Sem}}(J)||_{L^2} \tag{5.4}$$

It thus becomes possible to use our short *semantic* signature of size $M = 8$ in order to retrieve pCLE videos that are the closest to a video query according to the *semantic* distance d_{Sem}. In Section 5.6 we demonstrate that, in terms of correlation with the perceived visual similarity, the retrieval performance of the *semantic* distance d_{Sem} is comparable to that of the visual distance d_{Vis}.

In order to provide the endoscopists with a qualitative visualization of *semantic* signatures, we provide an intuitive representation of any *semantic* signature using a star plot of M radii, as shown in Fig. 5.5. Given a video I and the j^{th} semantic concept, we normalize the *semantic weight* s_j^I into $(s_j^I - \min\{s_j^J, J \in D^{train}\})/(\max\{s_j^J, J \in D^{train}\} - \min\{s_j^J, J \in D^{train}\})$ in order to obtain the coordinate value of I along the j^{th} radius of the star plot. For example, in Fig. 5.6 the star plots represent, from some tested videos, the visual-word-based *semantic* signatures that have been learned from annotated training videos, such as the ones shown in Fig. 5.2.

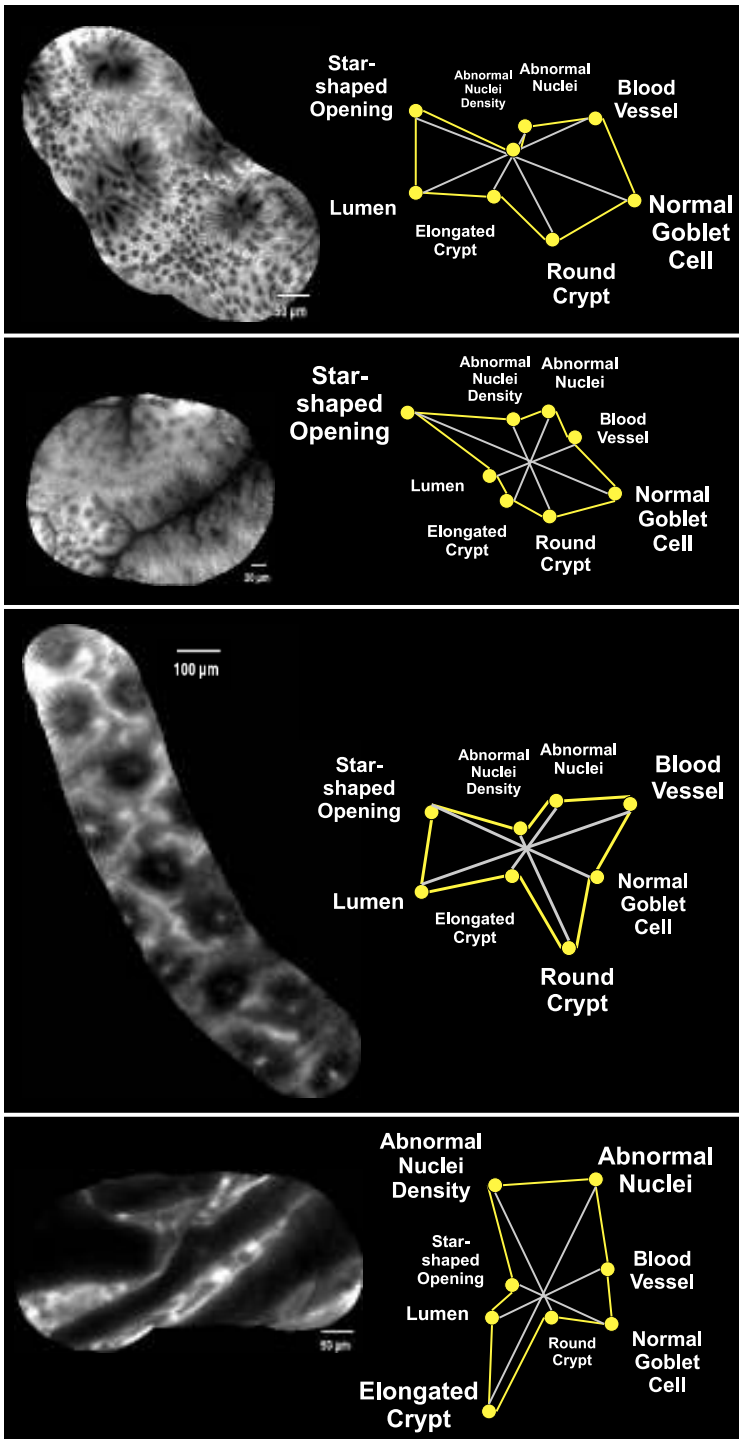


Figure 5.6: Examples of tested pCLE videos, represented by mosaic images, and visualization of their learned *semantic* signatures using the star plot, as explained in Fig. 5.5. The font size of each written semantic concept is proportional to the value of the concept coordinate in the star plot. Underlined concepts are those which were annotated as present in the semantic ground truth. From top to bottom, the first three mosaics show non-neoplastic (i.e. non-malignant) colonic polyps and the fourth mosaic shows a neoplastic (i.e. malignant) colonic polyp.

5.5 Distance Learning from Perceived Similarity

Similarity distance learning has been investigated by recent studies to improve classification or recognition methods. Yang et al. [Yang 10] proposed a boosted distance metric learning method that projects images into a Hamming space where each dimension corresponds to the output of a weak classifier. Weinberger and Saul [Weinberger 09] explored convex optimizations to learn a Mahalanobis transformation such that distances between nearby images are shrunk if the images belong to the same class and expanded otherwise. At the level of image descriptors, Philbin et al. [Philbin 10] have a similar approach that transforms the description vectors into a space where the clustering step more likely assigns matching descriptors to the same visual word and non-matching descriptors to different visual words. In order to model the perceived visual similarity between digital mammograms, El Naqa et al. [El-Naqa 04] proposed a hierarchal similarity learning approach, based on neural networks and support vector machines, which allows the incorporation of relevance feedback.

In order to improve the relevance of pCLE retrieval, our objective is to shorten the distances between *very similar* videos and to enlarge the distances between non-*very similar* videos. As the approach of Philbin et al. [Philbin 10] is closer to our pairwise visual similarity ground truth, we propose a generic distance learning technique inspired from their method. We aim at finding a linear transformation matrix W which maps given video signatures to new signatures that better discriminate *very similar* video pairs from the other video pairs. We thus consider two groups: D_+ is the set of N_+ training video couples that have been scored with $+2$ and D_- is the set of N_- training video couples that have been scored with $+1$, -1 or -2. We optimize the transformation W by minimizing the following margin-based cost function f:

$$\begin{aligned} f(W,\beta,\gamma) = \quad & \frac{1}{N+} \sum_{(I,J)\in D_+} g(\beta - d(W\ \mathscr{S}(I), W\ \mathscr{S}(J))) \\ & +\gamma\ \frac{1}{N_-} \sum_{(I,J)\in D_-} g(d(W\ \mathscr{S}(I), W\ \mathscr{S}(J)) - \beta) \end{aligned} \tag{5.5}$$

where $\mathscr{S}(I)$ is the signature of the video I, $d(.,.)$ is the chosen distance between the video signatures, e.g. L^2 or χ^2, and $g(z) = log(1 + e^{-z})$ is the logistic-loss function. The cost function f has the three following parameters: the transformation matrix W, the margin β and the constant parameter γ that potentially penalizes either non-*very similar* nearby videos or *very similar* remote videos. We could optimize f with respect to all 3 parameters, but this would make the search for the optimum more sensitive to local minima. We therefore decide to fix the value of the margin β using an intuitive heuristic: we take as a relevant value for β the threshold on the distances between video signatures that maximizes the classification accuracy between D_+ and D_-. All possible values of the parameter γ

are then discretized into a finite number of values, at which the cost function f is optimized according to W. As long as the distance $d(.,.)$ is differentiable, f can be differentiated with respect to W. Given a pCLE video I, its signature $\mathscr{S}(I)$ of size X is mapped to the transformed signature $W^{opt}\ \mathscr{S}(I)$, where W^{opt} is the optimized transformation matrix of size $X \times X$. The learned similarity distance between two pCLE videos I and J is then defined as:

$$d^{learn}(I,J) = d(W^{opt}\mathscr{S}(I), W^{opt}\mathscr{S}(J))) \tag{5.6}$$

The optimal value of γ, determined using cross-validation, is the one that maximizes the Pearson correlation coefficient between the learned similarity distance d^{learn} and the perceived similarity.

The application of this generic distance learning scheme to the *semantic* signatures of size $X = 8$ is straightforward: the transformation matrix W is of size $X \times X = 64$, $\mathscr{S} = \mathscr{S}_{\text{Sem}}$, the intuitive distance is $d(x,y) = ||x-y||_{L^2}$. Our experiments with cross-validation led to $\gamma = 10$.

However, for the application on the *visual* signatures of size $X = 100$, $\mathscr{S} = \mathscr{S}_{\text{Vis}}$ and the $X \times X = 10,000$ coefficients of the transformation matrix W should be positive in order to maintain the positiveness of visual word frequencies. Besides, as our sample size is relatively small, there is a risk of overfitting if all the $10,000$ coefficients of W are involved in the optimization process. For this reason, we only consider the optimization of diagonal matrices W, which amounts to optimize $K = 100$ visual word weights. Finally, the χ^2 pseudo-distance, initially used between visual word signatures, is an intuitive distance $d(.,.)$ between the transformed visual word signatures which should be L^1-normalized before χ^2 measures are performed:

$$d(W\ \mathscr{S}_{\text{Vis}}(I), W\ \mathscr{S}_{\text{Vis}}(J)) = \chi^2\left(\frac{W\ \mathscr{S}_{\text{Vis}}(I)}{||W\ \mathscr{S}_{\text{Vis}}(I)||_{L^1}}, \frac{W\ \mathscr{S}_{\text{Vis}}(J)}{||W\ \mathscr{S}_{\text{Vis}}(J)||_{L^1}}\right) \tag{5.7}$$

Due to the choice of the χ^2 pseudo-distance, the differentiation of the cost function f with respect to W is less straightforward but feasible. We also tried the L^2 distance for the distance $d(.,.)$ but we did not retain it because the results were not as good as with the χ^2 pseudo-distance. Our experiments with cross-validation for the *visual* signatures also led to $\gamma = 10$.

5.6 Evaluation and Results

5.6.1 Cross-validation

In order to exclude any learning bias or patient-related bias, we used $m \times q$-fold cross-validation, i.e. m random partitions of the database into q subsets, such that: each subset contains approximately the same number of patients, and all the videos of a same patient are in the same subset. Each of these subsets is successively the testing set and the union of the $q-1$ others is the training set. Given our sparse ground truth for perceived similarity, q must be small enough in order to

have enough similarity scores in the testing set, and large enough to ensure enough similarity scores in the training set. For our experiments, we performed $m = 30$ random partitions of our pCLE video database into $q = 3$ subsets. When computing any performance indicator, we will consider as a robust indicator value the median of all the indicator values computed with cross-validation.

5.6.2 Evaluation of Semantic Concept Extraction

Methodology

In order to evaluate, from the semantic point of view, our visual-word-based *semantic* extraction method, we propose to measure the performance of each of the $M = 8$ *semantic weights* contained in the *semantic* signature, using classification. For the j^{th} semantic concept, we compute a ROC curve that shows the matching performance of the learned *semantic weight* s_j with respect to the semantic ground truth c_j. The obtained ROC curves reflect how well the presence of semantic concepts can be learned from the visual words.

Results

From the semantic point of view, the performance of the *semantic* signature can be appreciated in the ROC curves shown in Fig. 5.7. The semantic concepts, from the best classified to the worst classified, are: *elongated crypt*, *round crypt*, *abnormal nuclei density*, *normal goblet cell*, *abnormal nuclei*, *lumen*, *blood vessel* and *star-shaped opening*. The fact that the concept *elongated crypt* is very well classified shows that the visual words clearly express whether this concept is present or not in pCLE videos. As the presence of elongated crypts in a pCLE video is a typical criterion of malignancy for the endoscopists, we deduce that *semantic* signatures could be successfully used for pCLE classification between malignant and non-malignant colonic polyps. Although the concepts *blood vessel* and *star-shaped opening* are poorly classified, they contribute to the clinical relevance of the visual-word-based *semantic* retrieval because their ROC curves are above the diagonal. Indeed, we will show in the next sections that these concepts act as "weak classifiers" for boosting similarity distance learning.

As the semantic classification deriving from *semantic* signatures is based on a rather intuitive Fisher-based linear method, it is worth to be compared with more sophisticated classification method, such as Support Vector Machines (SVM). We thus test the classification performance of a linear SVM and a non-linear SVM based on radial basis functions, which we feed with the visual word signatures. The resulting ROC curves are shown in Figs. 5.8 and 5.9. According to the areas under the ROC curves shown in Table 5.2, most of the ROC curves obtained with the linear SVM and with the non-linear SVM are statistically worse than those obtained with our intuitive method, and none of them are statistically better. These comparison results demonstrate the relevance of our intuitive Fisher-based method in terms of semantic classification, and thus the relevance of the *semantic* signatures.

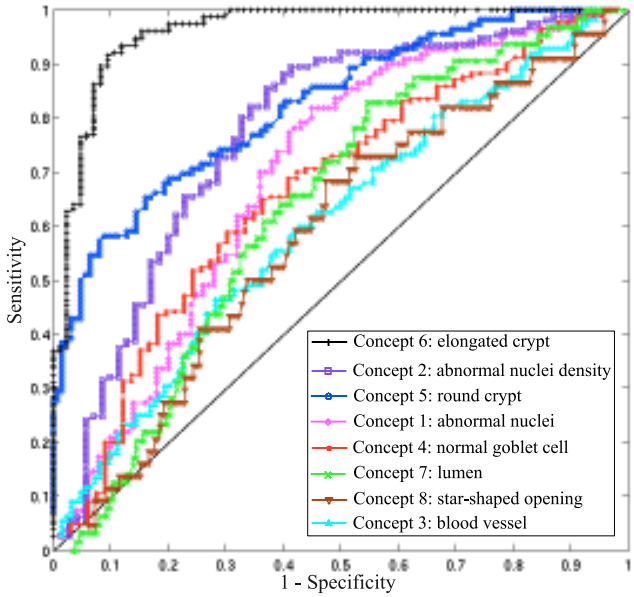


Figure 5.7: ROC curves showing, for each semantic concept, the classification performance of the *semantic* signature $\mathscr{S}_{\mathbf{Sem}}$. Each ROC curve associated with a concept c_j is the median of the ROC curves computed with 30×3 cross-validation by thresholding on the *semantic weight* s_j.

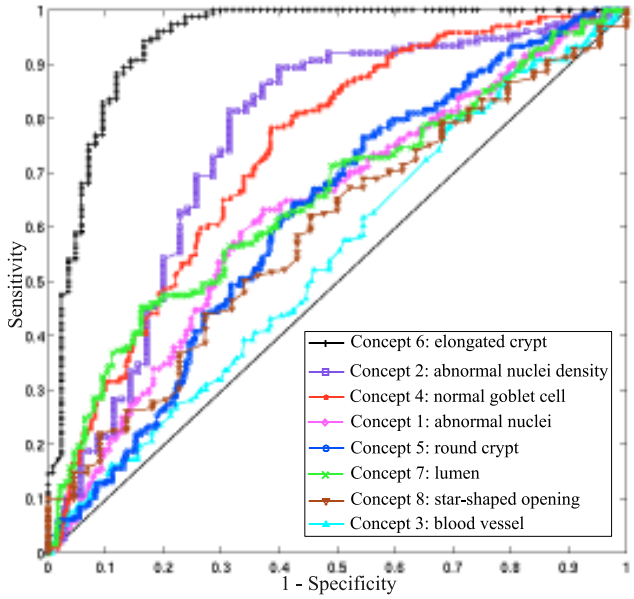


Figure 5.8: ROC curves showing the semantic classification performed by a non-linear SVM (based on radial basis functions) fed with the visual word signatures. Each ROC curve associated with a concept c_j is the median of the ROC curves computed with 30 × 3 cross-validation.

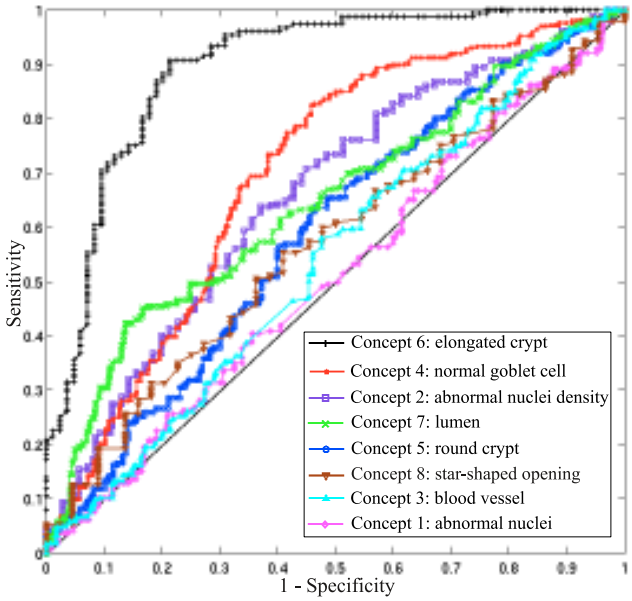


Figure 5.9: ROC curves showing the semantic classification performed by a linear SVM fed with the visual word signatures. Each ROC curve associated with a concept c_j is the median of the ROC curves computed with 30×3 cross-validation.

Semantic concept	**AUC Non-linear SVM**	**AUC Linear SVM**	**AUC Our method**
c1. abnormal nuclei	51.4 %	62.9 %	75.6 %
c2. abnormal nuclei density	65.8 %	75.7 %	81.8 %
c3. blood vessel	53.6 %	55.0 %	66.2 %
c4. normal goblet cell	69.8 %	73.6 %	71.6 %
c5. round crypt	58.9 %	62.0 %	86.3 %
c6. elongated crypt	89.6 %	94.2 %	96.7 %
c7. lumen	64.7 %	65.5 %	68.6 %
c8. star-shaped opening	57.0 %	59.4 %	62.8 %

Table 5.2: Area under the ROC curves (AUC) for each classification method according to each semantic concept. The corresponding ROC curves are shown in Figs. 5.9, 5.8 and 5.7. According to the AUC comparisons, our proposed intuitive Fisher-based method outperforms the linear SVM method (with statistical significance for all semantic concepts except $c4$). Besides, our method outperforms the non-linear SVM method for all semantic concepts except $c4$ (for which the performance is not statistically worse) and statistical significance is demonstrated for the concepts $c1$, $c2$, $c3$ and $c5$.

5.6.3 Retrieval Evaluation Tools

Methodology

Standard recall curves are a common means of evaluating retrieval performance. However, because of the sparsity of our perceived similarity ground truth, it is not possible to compute them in our case. As an alternative, we define *sparse recall* curves. At a fixed number k of nearest neighbors, we define the *sparse recall* value of a retrieval method as the percentage of L-scored video couples, with $L = +2$ (or $L >= 1$), for which one of the two videos has been retrieved among the k nearest neighbors of the other video. The resulting *sparse recall* curve shows the ability of the retrieval method to extract, among the first nearest neighbors, videos that are perceived as *very similar* to the video query.

The evaluation of a retrieval method against perceived similarity ground truth can be qualitatively illustrated by four superimposed histograms H_L, $L \in \{-2, -1, +1, +2\}$. H_L is defined as the histogram of the similarity distances which were computed by the retrieval method in the restricted domain of all L-scored video couples, where L is one of the four Likert points: *very dissimilar* ($L = -2$), *rather dissimilar* ($L = -1$), *rather similar* ($L = +1$) and *very similar* ($L = +2$). The more separated these four histograms are, the more likely the distance computed by the retrieval method will be correlated with perceived similarity ground truth. We use the Bhattacharyya distance as a separability measure between each pair of histograms.

Possible indicators of the correlation between the distance computed by a retrieval method and the perceived similarity ground truth are Pearson correlation π, Spearman ρ and Kendall τ. Compared to Pearson π which measures linear dependence based on the data values, Spearman ρ and Kendall τ are better adapted to the psychometric Likert scale because they measure monotone dependence based on the data ranks [Barnett 91]. Kendall τ is less commonly used than Spearman ρ but its interpretation in terms of probabilities is more intuitive. To assess statistical significance for the comparison between two correlation coefficients associated with two retrieval methods, we have to perform the adequate statistical test. First, ground-truth data lying on the four-point Likert scale can obviously not be characterized by a normal distribution. Data ranks should be used instead of data values. Second, the rank correlation coefficients measured for two methods are themselves correlated because they both depend on the same ground-truth data. For these reasons, we decide to perform Steiger's Z-tests, as recommended by Meng et al. [Meng 92], and we apply it to Kendall τ. We refer the reader to the Appendix A for a detailed description of the Steiger's Z-test.

Results

For our experiments, we compared the retrieval performances of "Dense-Sift" with those of "HH-Sift", "Haralick" and "Textons" presented in Section 5.3 which are considered as state of the art in CBIR. We call "Semantic" the visual-word-based

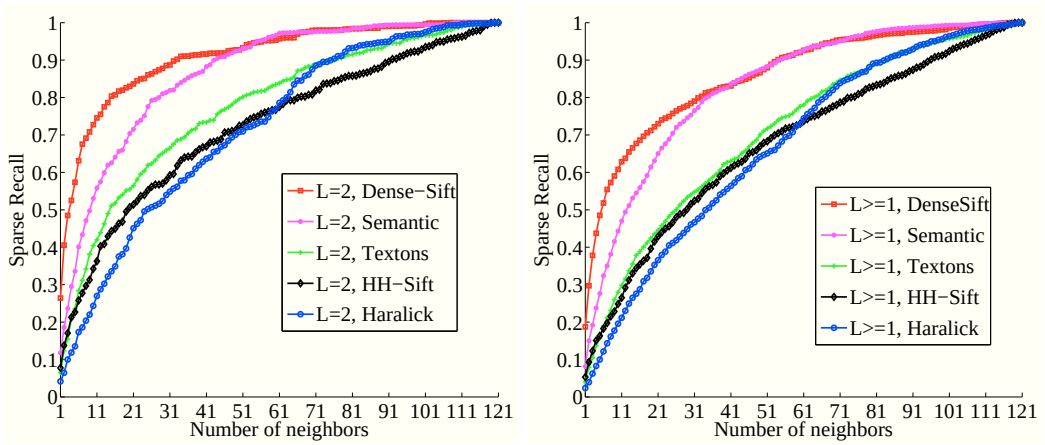


Figure 5.10: *Sparse recall* curves associated with the retrieval methods in L-scored domains where $L = +2$ (left) or $L >= 1$ (right).

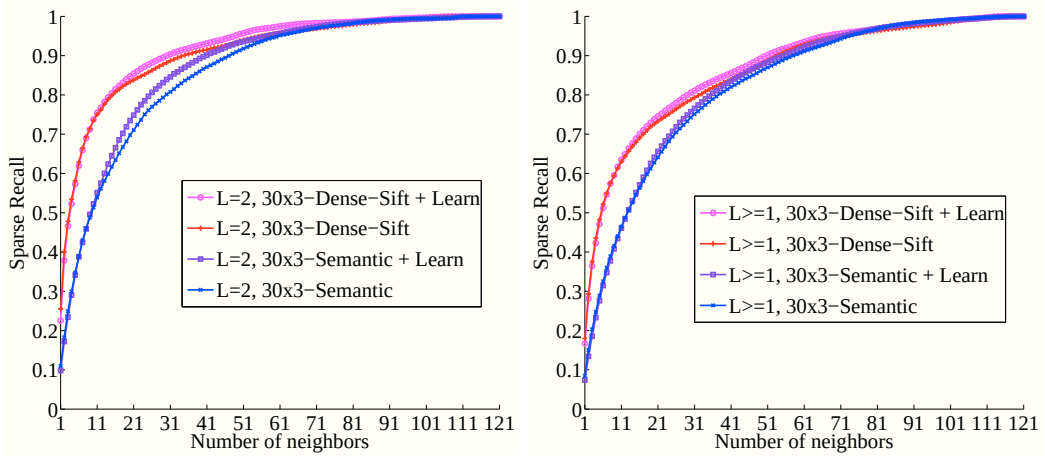


Figure 5.11: *Sparse recall* curves, before and after distance learning using cross-validation, in L-scored domains where $L = +2$ (left) or $L >= 1$ (right). Each *sparse recall* curve is the median of the *sparse recall* curves computed with 30×3 cross-validation.

semantic retrieval method, "30x3-Semantic" the same method with 30×3 cross-validation and "30x3-Dense-Sift" the "Dense-Sift" with 30×3 cross-validation. "30x3-Semantic+Learn" (resp. "30x3-Semantic+Learn") is the "30x3-Semantic" method (resp. "30x3-Dense-Sift+Learn" method) improved with distance learning.

In terms of *sparse recall* performances, we observe in Figs. 5.10 and 5.11 that the retrieval methods from best to worst are: "Dense-Sift+Learn", "Dense-Sift", "Semantic+Learn", "Semantic", "Textons", "HH-Sift" and "Haralick". In particular, perceived similarity distance learning allows to slightly improve recall performance. The fact that "Dense-Sift" outperforms "Semantic" before and after distance learn-

Bhattacharyya distance between Hist(L) and Hist(L')	$L=+2$ $L'=+1$	$L=+2$ $L'=-1$	$L=+2$ $L'=-2$	$L=+1$ $L'=-1$	$L=+1$ $L'=-2$	$L=-1$ $L'=-2$
10x3-Sem+Learn	0.024	0.175	0.468	0.078	0.294	0.072
10x3-Sem	0.018	0.145	0.441	0.071	**0.299**	**0.075**
10x3-DS+Learn	**0.036**	**0.236**	**0.500**	**0.087**	0.254	0.047
10x3-DS	0.030	0.205	0.412	0.084	0.219	0.036
Semantic (Sem)	0.046	0.200	**0.571**	0.090	**0.352**	**0.102**
Dense-Sift (DS)	**0.051**	**0.257**	0.519	**0.096**	0.251	0.051
Textons	0.030	0.152	0.193	0.067	0.095	0.023
Haralick	0.042	0.089	0.206	0.038	0.125	0.048
HH-Sift	0.037	0.098	0.102	0.047	0.042	0.027

Table 5.3: Measures of separability, using Bhattacharyya distance, between the four L-scored histograms H_L shown in Figs 5.12 and 5.13 for each retrieval method. For the retrieval methods using 30×3 cross-validation, we computed the median of the Bhattacharyya distances.

ing might be explained by the small size of the *semantic* signatures ($M = 8$) with respect to the larger size of the *visual* signatures ($K = 100$): *semantic* signatures might be too short to discriminate *very similar* video pairs with the same performance as *visual* signatures.

On the superimposed histograms shown in Figs. 5.12 and 5.13, we observe qualitatively that "Dense-Sift" and "Semantic" globally better separate the four histograms than "HH-Sift", "Haralick" and "Textons", and that perceived similarity distance learning allows to better separate the histogram H_{+2} from the other histograms. These observations are quantitatively confirmed by the Bhattacharyya distances shown in Table 5.3. The correlation results shown in Tables 5.4 and 5.5 also confirm these findings and demonstrate that, with statistical significance, the similarity distances computed by "Dense-Sift" and "Semantic" are better correlated with the perceived similarity than the similarity distances computed by "HH-Sift", "Haralick" and "Textons". Besides, with statistical significance, the learned similarity distances are better correlated with the perceived similarity than the original distances. These results also show that the correlation performance of "30x3-Semantic+Learn" (resp. "30x3-Semantic") is comparable to that of "30x3-Dense-Sift+Learn" (resp. 30x3-Dense-Sift"), as their difference is not statistically significant.

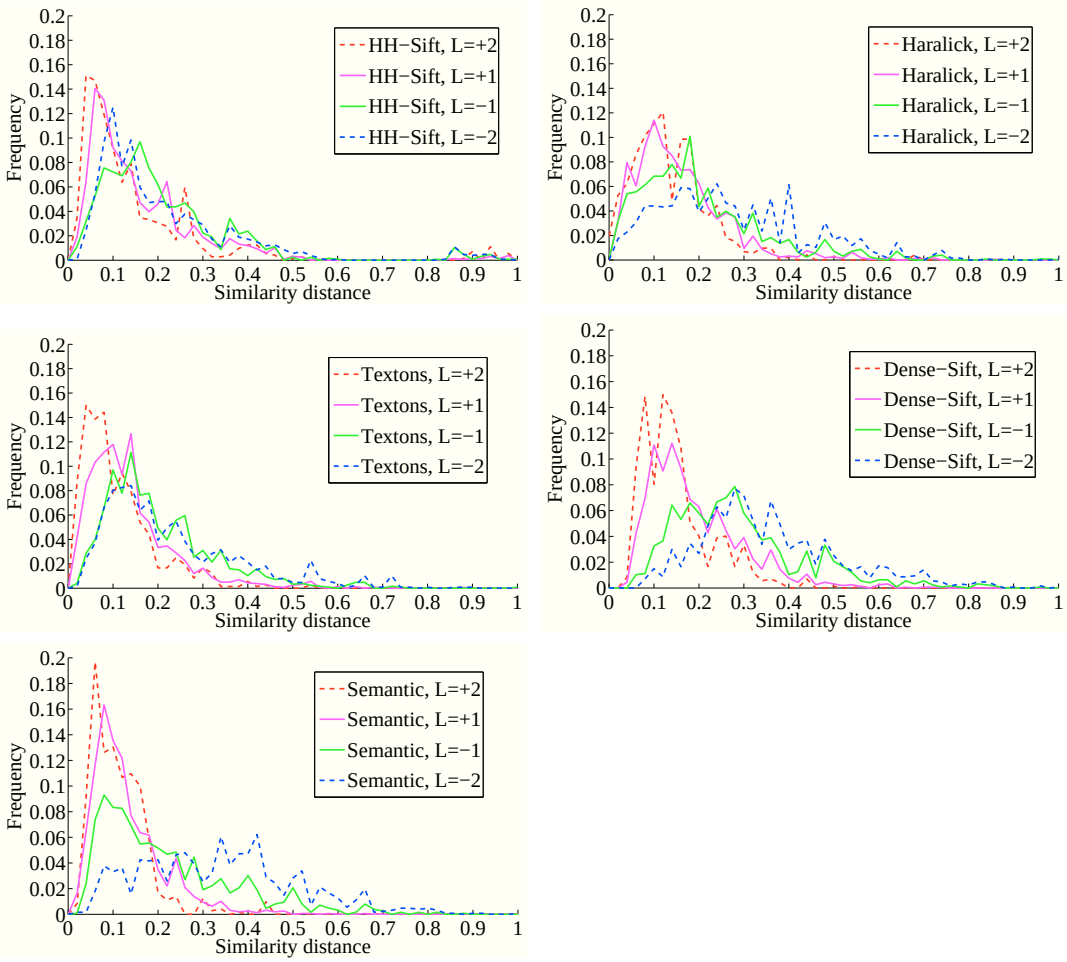


Figure 5.12: Superimposed histograms H_L of the similarity distances in each L-scored domain. From top left to bottom right: "HH-Sift" method, "Haralick" method, "Textons" method, "Dense-Sift" method, "Semantic" method.

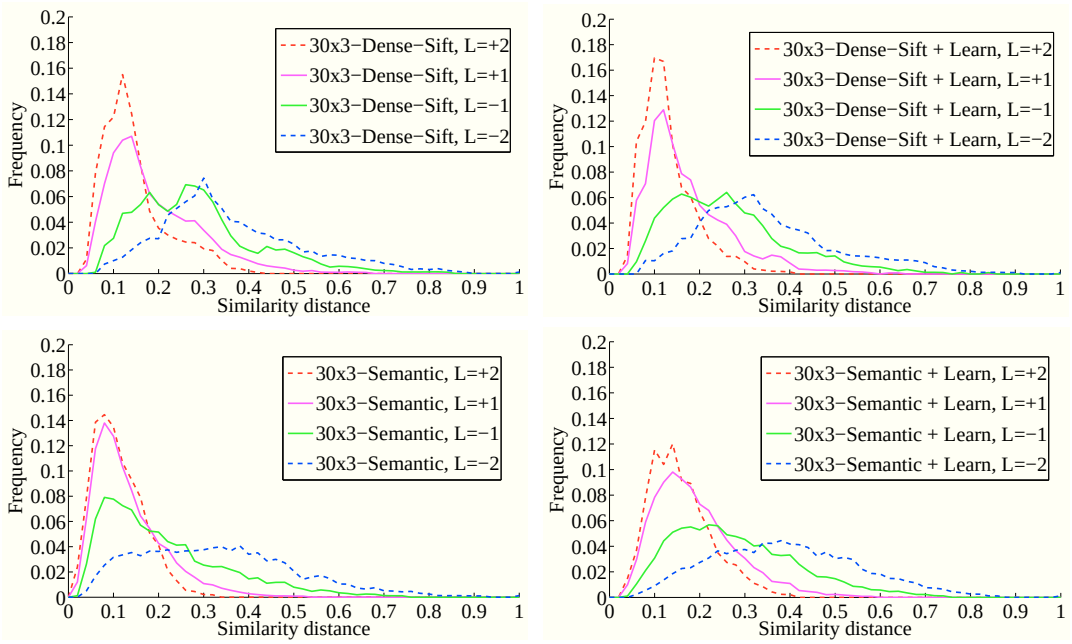


Figure 5.13: Superimposed histograms H_L of the similarity distances in each L-scored domain. On the top: "30x3-Dense-Sift" method (left) and "30x3-Dense-Sift+Learn" method (right). **On the bottom:** "30x3-Semantic" method (left) and "30x3-Semantic+Learn" method (right). Each histogram is the median of the histograms computed with 30×3 cross-validation.

Retrieval method	**M1** Sem	**M2** DS	**M3** Textons	**M4** Haralick	**M5** HH-Sift
Pearson π	54.6 %	51.6 %	35.3 %	35.4 %	15.8 %
Spearman ρ	55.3 %	55.7 %	38.2 %	34.5 %	22.8 %
Kendall τ	49.4 %	**50.0** %	34.1 %	30.4 %	20.0 %
Steiger's Z-test on τ; p-value	**> M3,M4 > M5** $p < 10^{-45}$	**> M3,M4 > M5** $p < 10^{-60}$	**> M4 > M5** $p < 10^{-4}$	**> M5** $p < 10^{-15}$	
	$\sim$ M2 $p = 0.486$				

Table 5.4: Indicators of correlation between the similarity distance computed by the retrieval methods and the ground truth. > M indicates that the improvement from method **M** is statistically significant, **$\sim$ M** indicates that it is not.

Retrieval method	**M1"** 10x3-Sem+Learn	**M1'** 10x3-Sem	**M2"** 10x3-DS+Learn	**M2'** 10x3-DS
Pearson π	55.7 %	53.3 %	53.4 %	51.4 %
σ	0.3 %	0.2 %	0.2 %	0.2 %
Spearman ρ	56.6 %	53.8 %	58.2 %	55.5 %
σ	0.3 %	0.2 %	0.2 %	0.3 %
Kendall τ	50.9 %	48.1 %	**52.4** %	49.8 %
σ	0.3 %	0.2 %	0.2 %	0.2 %
Steiger's Z-test on τ; p-value	**> M1'** $p = 0.022$		**> M1'** **> M2'** $p < 0.003$	
	$\sim$ M2",M2' $p > 0.05$	**$\sim$ M2'** $p = 0.163$	**$\sim$ M1"** $p > 0.05$	

Table 5.5: Indicators of correlation between the similarity distance computed by the retrieval methods and the ground truth. After performing 30×3 cross-validation, we compute and show the median of correlation coefficients. The standard deviation σ of each correlation estimator can be computed from the standard deviation of the n samples $\sigma_{samples} = \sqrt{n-1}\sigma$. We also show the median of p-values when comparing two retrieval methods using 30×3 cross-validation. $>$ **M** indicates that the improvement from method **M** is statistically significant, $\sim$ **M** indicates that it is not.

5.6.4 Discussion

Looking at the *sparse recall* curves, although the results based on *semantic* signatures are not as good as those based on *visual* signatures, the curve of *semantic* signatures is much closer to the curve of *visual* signatures than the curves of state-of-the-art methods. We can therefore be rather confident in the fact that the *semantic* signatures are informative. *Sparse recall* is only a means to evaluate the relevance of the *semantic* signatures. Indeed, we want to base the retrieval of pCLE videos on visual content and not on semantic annotations, otherwise the retrieval system might retrieve videos that are semantically related but not similar in appearance, in which case the physician might lose trust in the retrieval system. In order to ensure both the higher recall of the visual word retrieval method after distance learning, and the clinical relevance of the semantic information contained in the *semantic* signature, we propose a pCLE retrieval system where the most similar videos are extracted using the "Dense-Sift+Learn" method, and where the star plots representing *semantic* signatures are displayed. Figs. 5.14 and 5.15 shows some typical results of our pCLE retrieval system with 5 nearest neighbors, with the added semantic ground truth represented by underlined concepts. In clinical practice, the semantic ground truth is not known for the video query, but in these retrieval examples it is disclosed for illustration purposes. The extracted pCLE videos, represented as mosaic images, look quite similar in appearance to the query, the first neighbor being more visually similar than the last one. On each star plot, the font size of each written semantic concept is proportional to the normalized value of its *semantic weight*. Semantic concepts written in large characters may or may not be in agreement with the underlined concepts present in the ground truth. Most importantly, if for a given pCLE video, the semantic ground truth is very different from the estimated *semantic* signature, then the difficulty to interpret the video for diagnosis purpose might be high, because visual content is not correlated with semantic annotations. Our visual-word-based *semantic* signature would thus have the potential to distinguish ambiguous from non-ambiguous pCLE videos. The remaining disagreements between the learned semantic information and the semantic ground truth show that, even though we have achieved encouraging results in extracting semantics from visual words, further investigations are still needed to bridge the semantic gap between low-level visual features and high-level clinical knowledge.

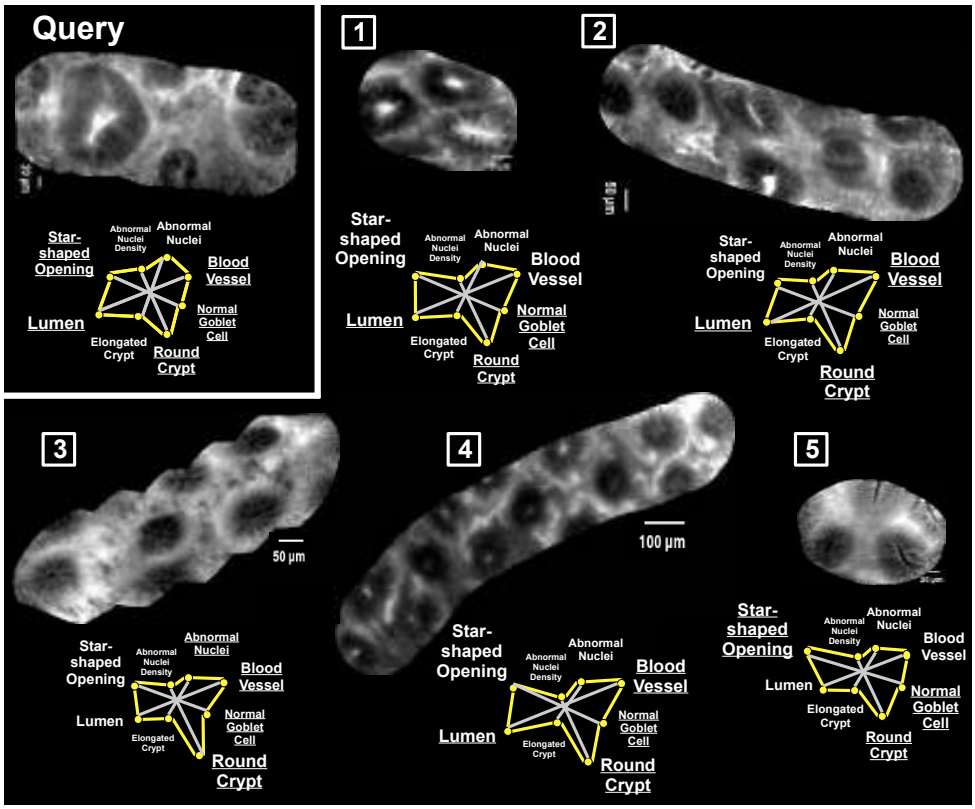


Figure 5.14: Examples of pCLE retrieval results from a non-neoplastic video query. The 5 most similar videos are retrieved by "30x3-Dense-Sift+Learn" method. For each video, the star plot representation of its *semantic* signature is provided. The font size of each written semantic concept is proportional to the value of the concept coordinate in the star plot. Underlined concepts are those which were annotated as present in the semantic ground truth. In practice, the semantic ground truth is not known for the video query, but it is disclosed here for illustration purposes. For illustration purposes, videos are represented by mosaic images.

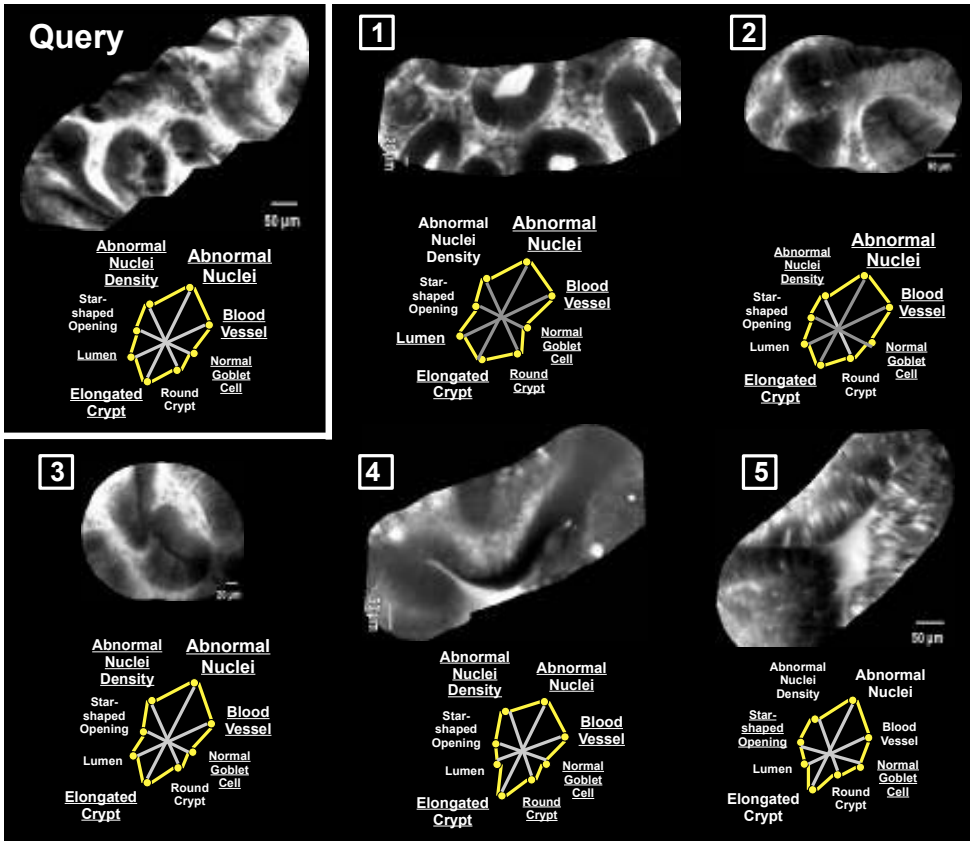


Figure 5.15: Examples of pCLE retrieval results from a neoplastic video query. The 5 most similar videos are retrieved by "30x3-Dense-Sift+Learn" method. For each video, the star plot representation of its *semantic* signature is provided. For illustration purposes, videos are represented by mosaic images.

5.7 Conclusion

The pCLE retrieval system proposed in this study provides the endoscopists with clinically relevant information, both visual and semantic, that should be easily interpretable to make an informed pCLE diagnosis. Our main contributions in this chapter are: (1) a Fisher-based method that builds short visual-word-based *semantic* signatures, (2) an intuitive representation of these *semantic* signatures using star plots, (3) the creation of an online tool to generate a relevant ground truth for visual similarity perceived by multiple endoscopists between pCLE videos, (4) a method for distance learning from perceived visual similarity to improve retrieval relevance, and (5) the implementation of several tools to evaluate retrieval methods, such as correlation measures and *sparse recall* curves. Moreover, this proposed methodology could be applied to other medical or non-medical databases, as long as ground-truth data is available.

Despite our relatively small pCLE database and despite the sparsity of the perceived similarity ground truth, our evaluation experiments show that the visual-word-based *semantic* signatures extract, from low-level visual features, a higher-level clinical knowledge which is consistent with respect to perceived similarity. Possible disagreements between the semantic estimation, based on visual features, and the semantic ground truth could be investigated in order to estimate the interpretation difficulty of pCLE videos, which we explored in Chapter 4 only based on visual words. Future work will focus on more sophisticated methods to learn jointly visual and semantic similarity. Our long-term objective is the clinical evaluation of our visual-semantic retrieval system to see whether it could help the endoscopists in making more accurate pCLE diagnosis.

Chapter 6

Conclusions

Table of Contents

French summary

Notre but à travers ce manuscrit a été de démontrer comment le diagnostic endomicroscopique in vivo *des cancers gastro-intestinaux peut être facilité par un système de reconnaissance d'images par le contenu, ajusté et combiné avec l'apprentissage de connaissances cliniques de plus haut niveau. Ce système de reconnaissance, complété par l'estimation de caractéristiques non visuelles telles que la difficulté d'interprétation et les concepts sémantiques, constitue notre atlas intelligent. Les applications cliniques de cet atlas intelligent couvrent l'aide au diagnostic, la formation et le partage des connaissances en endomicroscopie.*

Les principales perspectives que avons identifiées pour la suite de cette thèse sont la validation clinique, la généralisation à d'autres organes et à d'autres applications, une meilleure utilisation des informations spatio-temporelles, et l'élaboration d'un système multimodal de reconnaissance capable d'inclure, en plus des images endomicroscopiques, des images endoscopiques et histologiques, ainsi que des métadonnées textuelles.

6.1 Contributions and Clinical Applications

Our goal throughout this manuscript has been to show how the *in vivo* endomicroscopy diagnosis of gastrointestinal cancers can be supported by a content-based image retrieval approach that is adjusted and combined with the learning of higher-level clinical knowledge. The resulting pCLE retrieval system, augmented with the estimation of non-visual features such as interpretation difficulty and semantic concepts, is our proposed "Smart Atlas". The clinical applications of the "Smart Atlas" include pCLE diagnosis support, training support and knowledge sharing.

Our successive contributions have developed generic and objective tools, ranging from low-level to high-level feature extraction tools, in order to guide the endoscopists in their subjective interpretation of pCLE video sequences. Although these retrieval tools could be easily extended for other medical or non-medical

applications, we have presented in this thesis, their concrete applications to the early diagnosis of gastrointestinal cancers, using two different pCLE databases: the *Colonic Polyps* and the *Barrett's Esophagus*. This has resulted in several clinical publications [André 10b, André 10b, André 11a]. Our evaluation methods for measuring retrieval performance have also evolved, from indirect evaluations using classification in [André 09a, André 10c, André 11e], to direct evaluations using a perceived similarity ground truth in [André 11d, André 11c]. Despite the relatively small size of the annotated pCLE databases and despite the sparsity of our perceived similarity ground truth, these evaluation results are quite encouraging and constitute a robust proof-of-concept before further clinical evaluations.

Our first main contribution has been the adjustment of the bag-of-visual-words method for the retrieval of single pCLE images in [André 09a], and of full pCLE videos represented as a set of mosaic images in [André 09b, André 10c]. Most importantly, the choice of a dense image descriptor and the manipulation of "implicit mosaics" has allowed us to efficiently build one short visual word signature per pCLE video. As these *visual* signatures adequately represent the pCLE videos, the use of a standard distance between them provides relevant retrieval results which can be qualitatively appreciated. We have also investigated several orthogonal methods, including multi-scale description, visual word discriminative power and spatial relationships between local features. Unfortunately, probably due to the relatively small size of the database, we did not manage to demonstrate a significant impact of these methods on the retrieval performances. We thus consider these original methods as relevant proofs of concept until we have larger databases. Using leave-one-patient-out cross-validation, we have demonstrated that, on two different pCLE databases, our pCLE retrieval method outperforms several state-of-the-art methods CBIR, with statistical significance. In [André 11a], we have also shown that our pCLE retrieval method is comparable to the offline pCLE diagnosis of expert endoscopists.

Another important contribution of this thesis has been the automated estimation, based on retrieval results, of the interpretation difficulty attached to a pCLE video. In [André 10a], a significant relationship has been shown between the estimated difficulty and the diagnosis difficulty experienced by multiple endoscopists. Such a difficulty estimator can be used for the development of a self-training simulator featuring difficulty level selection in order to help the endoscopists in shortening their learning curve in pCLE diagnosis. Difficulty estimation can also be used to complement the outputs of video retrieval by indicating a confidence level associated to the query.

Regarding retrieval evaluation, we have succeeded in moving beyond a critical issue which was the lack of an objective ground truth for CBIR. Indeed, we have developed an online survey tool in [André 11d] which allows multiple experts in pCLE to individually evaluate the visual similarity that they perceive between pCLE videos. Thanks to the resulting ground truth for perceived similarity, direct evaluations and comparisons of retrieval performances have then been possible, by measuring the correlation between retrieval distance and perceived similarity, or by

generating what we have defined as "sparse recall" curves.

Having a ground truth for perceived similarity has also allowed, in [André 11d], us to learn an adequate visual similarity distance between pCLE videos, which we have proved to be more accurate, with statistical significance, than the original retrieval distance. This learned visual similarity distance can be used to define the "typicality" of a pCLE video based on its average distance to a given number of neighboring pCLE videos. The relevance of typicality estimation highly depends on the representativity of the training database, i.e. on how well the variability in the appearance of typical pathologies is represented by training pCLE videos. Furthermore, clustering pCLE videos according to the similarity distance would enable the hierarchical navigation of the endoscopists through the training video database at several levels of typicality.

Our last contribution, in [André 11c], has been the incorporation of clinical expertise for the visual-word-based learning of pCLE semantics. We have been able to transform the *visual* signatures into *semantic* signatures that reflect how much the presence of semantic concepts is expressed by the visual features in the videos. Thus, from low-level visual features, a higher-level clinical knowledge has been extracted, which is directly interpretable by the endoscopists and consistent with respect to perceived similarity. Ultimately, we have suggested to leverage the possible disagreements, between visual-word-based semantic estimation and semantic ground truth, for the automated estimation of the visual-semantic ambiguity in pCLE videos.

It is worth mentioning that the contributions of the "Smart Atlas" should help to answer important clinical questions which have been recently identified by the ASGE (American Society for Gastrointestinal Endoscopy) PIVI (Preservation and Incorporation of Valuable endoscopic Innovations) initiative on real-time endoscopic assessment of the histology of diminutive colorectal polyps [Rex 11]:

> *Certainly, a large set of broad-style questions about unknowns with feedback explaining the answers could be developed and made widely available to remote trainees over the web. Improved broadband access should allow such materials to incorporate video clips for an experience that more realistically approximates the real time decision making required to make optical biopsies. In particular,* a comprehensive image database with path correlation for "difficult to interpret" small polyps would be most important to develop in order to help trainees reduce the percentage of lesions which they can interpret only with low confidence. *Another suggestion that would require significant effort on the part of the endoscope manufacturers would be to develop* real time pop up image atlases to assist endoscopists with interpretation as they are performing the examination. *Similar to the tool available during reading of capsule endoscopy, this might become possible with future generations of endoscope systems. More technically feasible for the present would be the development and dissemination of very good posters to be hung in*

> *the endoscopy suites detailing the features of adenomatous and hyperplastic polyps using various imaging modalities. Whichever materials are ultimately utilized, these will be* helpful both in initial training and in ongoing reinforcement of interpretation skills. *The development of approved teaching materials by the ASGE with industry support, along with an approved and validated teaching program will clearly accelerate the acceptance of optical diagnosis in clinical practice.*

6.2 Perspectives

The following perspectives, regarding both research directions and clinical applications, are worth to be explored.

In order to ensure the clinical validation of the "Smart Atlas" tool, further clinical evaluations are required. Short-term evaluations will measure the impact of using the tool on self-training and on offline diagnosis. Long-term evaluations will measure its impact on online diagnosis, established during ongoing endoscopy, which implies more ergonomic and runtime constraints. In both cases, we suggest to use evaluation protocols based on the Second Reader Paradigm (SRP), which can be used in two ways SRP1 and SRP2. In the SRP1 way, the endoscopist establishes first a blinded diagnosis on a pCLE video of interest and then a second diagnosis on the same video but with the additional diagnostic information predicted by the "Smart Atlas". In case of disagreement with the "Smart Atlas", the endoscopist may wish to revise her judgment and see the retrieval results of the "Smart Atlas", as proposed by the SRP2 way. In the SRP2 way, a further diagnosis is established with the additional retrieval information of annotated pCLE videos, visually similar to the query, that are extracted by the "Smart Atlas". By comparing the diagnosis performances of the endoscopist, without and with the "Smart Atlas" support, we can measure whether the "Smart Atlas" help to improve the accuracy of pCLE diagnosis.

We plan to improve the performances of the "Smart Atlas", by training on pCLE video databases that are more representative of the atypical cases, and of the variability in the appearance of pathologies. To this purpose, we will include more pCLE videos to enlarge our two current databases, on colonic polyps and on Barrett's esophagus. We also plan to test the "Smart Atlas" on other other organs or other pathologies, as far as their appearance in pCLE videos contains sufficiently discriminative shape and texture information. It is the case of many medical applications of pCLE, for instance the duodenum or the Endoscopic Mucosal Resection (EMR).

Further steps towards bridging the semantic gap, between low-level visual features and high-level clinical knowledge, could be performed. For example, we will look at more sophisticated methods to jointly learn visual and semantic similarity distance between pCLE videos. Besides, some of the semantic concepts, which are not too scattered in pCLE images, could be automatically segmented in the videos,

from the analysis of their visual expressions. We may also consider a semantic query for the "Smart Atlas" that expects, from a given semantic concept, the extraction of multiple pCLE videos that are representative of the concept. In addition to visual queries, such semantic queries would create another way of navigating through the annotated pCLE databases, and thus constitute another training support.

We would like to investigate more advanced methods for spatio-temporal retrieval of pCLE videos. Although the full temporal information of pCLE videos is not exploited in mosaic images, it may be captured by $2D + t$ retrieval techniques. This would be an interesting approach to support the pCLE diagnosis of some pathologies that are characterized by discriminative motions within the observed region of interest. It is for example the case of chronic inflammations that are associated to the visible motion of blood cells in the vessels. Besides, in order to exploit the spatial relationship between local visual features in pCLE images, more advanced methods could also be explored. In particular, if a sufficiently large pCLE database is available, the spatial information contained in multi-scale co-occurrence matrices of visual words can be incorporated into the image descriptors, without the risk of overfitting.

Another important perspective would be multimodal information retrieval. Indeed, all pCLE databases could be enriched with multimodal information, including both image and text information. In addition to pCLE image sequences, the image information could be composed of the histological images and the endoscopic images, potentially with zoom endoscopy or narrow-band imaging. In addition to textual pCLE diagnoses, the text information could be composed of the histological diagnosis, with Paris or Kudo classification, the patient information such as age and history, and the location and the size of the suspicious area. Multimodal retrieval methods could then be developed by leveraging all these informations. However, a multimodal retrieval that uses both pCLE and endoscopic images would require a robust correspondence between the microscopic view of pCLE and the macroscopic view of endoscopy. Some research teams have already worked on this co-localization problem, from example Allain et al. [Allain 10] who proposed a system based on epipolar geometry for biopsy site re-targeting. Another issue of multimodal retrieval is that the heterogeneous database may be sparsely populated. Indeed, having an incomplete database implies the definition of a more complex distance on potentially missing attributes. Such a similarity distance could be for example derived from the random forests model, as the one suggested by Iglesias et al. [Iglesias 11], which is able to deal with missing data. The resulting multimodal retrieval system would be highly valuable for the cross-disciplinary understanding of cancer diagnosis.

An interesting application of pCLE retrieval would be, given a set a pathological categories, the extraction of the closest pCLE video in each of these categories. This would make the physician more aware of the potential ambiguity of the query video. In this application, we can expect large distances between the query and some extracted videos, for example if the query is a benign colonic polyp and one pathological category is adenocarcinoma. Consequently, pCLE retrieval in this case must be accurate not only for short-range distances but also for long-range

distances, which requires suitable distance learning techniques, such as manifold learning. Manifold learning would also be a quite relevant method to estimate the distances between pCLE videos of the same patient at different time points. Such a longitudinal approach should help in quantifying tumor evolution in a given patient, which would be useful for cancer surveillance.

Finally, the "Smart Atlas" tool could be a valuable aid not only to endoscopist users, but also to users from other medical disciplines, such as histopathology and surgery. Three different scenarios can be identified. First, the most straightforward extension of the "Smart Atlas" would be its application to other types of images, in particular the histological images. A "Smart Atlas" extended to histological data should support the pathologists in making a more accurate diagnosis. Second, a multimodal "Smart Atlas" based on both pCLE and histological data should also be useful to the pathologists: it would assist them in the interpretation of pCLE images that are new to them and that correspond to biopsies performed by the endoscopists. Third, during surgical procedures, where the surgeon does not want to take the responsibility of diagnosing the pathologies, the "Smart Atlas" could be an indirect support. Indeed, the recently created PERSEE project [PERSEE 10] would allow the surgeon and the pathologist to work in concert by connecting diagnosis and treatment with a telemedicine system. A "Smart Atlas" tool compatible with real-time conditions could thus support the pathologist in diagnosing pCLE data that are acquired by the surgeon during ongoing procedure. In return, the surgeon would be able to make more informed choices about how to treat cancer patients to ultimately improve their outcomes.

Appendix A: Statistical Analysis Methods

This appendix provides a description of the statistical tests used in the thesis, based on Sheskin's book "Handbook of Parametric and Nonparametric Statistical Procedures" [Sheskin 11].

McNemar's Test: is there a statistically significant difference?

The McNemar's test is a nonparametric statistical test employed to compare two experiments, in which each of n subjects contributes two scores, one for each experiment, on a dichotomous dependent variable, i.e. scores must fall within one of two mutually exclusive categories, c_1 and c_2.

For example: the n subjects are pCLE images, the two experiments are two classification methods, and the score of an image given a method is equal to 1 (resp. to 0) if the image has been correctly classified (resp. misclassified) by the method, according to the diagnosis ground-truth. The null hypothesis is that there no significant difference between the two experiments, i.e. classification methods.

Let $n_{1,2}$ be the number of subjects having a score c_1 for the first experiment, and a score c_2 for the second experiment. Let $n_{2,1}$ be the number of subjects having a score c_2 for the first experiment, and a score c_1 for the second experiment. Then the McNemar's test statistic is given by:

$$Q = \frac{(n_{1,2} - n_{2,1})^2}{n_{1,2} + n_{2,1}} \tag{1}$$

Under the null hypothesis, Q has a χ^2 distribution with one degree of freedom and the associated p-value provides the statistical relevance.

In our example and for a statistical significance level $\alpha = 5\%$: if p-value< 0.05, the null hypothesis is rejected, which means that the difference between the two classification methods is statistically significant.

Since the McNemar's test uses a continuous distribution to approximate the discrete binomial distribution, a correction for continuity is recommended with small sample size, usually when $n_{1,2} + n_{2,1} < 20$. The continuity-corrected version of the McNemar's test statistic is:

$$Q_{\text{corr}} = \frac{(|n_{1,2} - n_{2,1}| - 1)^2}{n_{1,2} + n_{2,1}} \tag{2}$$

For extremely small sample sizes, the McNemar's test should not be employed.

Two-Sided Z-Test Between Proportions: is there a statistically significant equivalence?

The two-sided Z-test between proportions allows to test if there a statistically significant equivalence between two experiments. It requires the assumption that the data are normally distributed. As pointed out by Jones et al. [Jones 96], absolute equivalence cannot be demonstrated: it is only possible to assert that the true difference is unlikely to be outside a predefined range of equivalence $[-\Delta, +\Delta]$.

For example: we have n observations on two variables X_1 and X_2, one for each experiment, with respective means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. The null hypothesis is that an absolute difference of at least Δ exists between the two experiments.

The two sided 95% confidence interval of the two-sided Z-test statistic is given by:

$$CI_{95} = (\mu_1 - \mu_2) \pm 1.96\ (\mu_1 - \mu_2)\ \sqrt{\frac{n-1}{\sigma_1^2 + \sigma_2^2}} \tag{3}$$

where 1.96 is the Z-value corresponding with 95% of the area under the standard normal distribution.

In our example and for a statistical significance level $\alpha = 5\%$: if the 95% confidence interval CI_{95}, centered on the observed difference $\mu_1 - \mu_2$, lies entirely within the predefined range of equivalence $[-\Delta, +\Delta]$, then there is a statistically significant equivalence between the two experiments.

Permutation test: is there a statistically significant correlation?

The permutation test is a nonparametric statistical test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under random rearrangements of the labels on the observed data points. The advantage of the permutation test is that it does not require any a priori assumption about the distribution of the data, as it generates all possible permutations of the data to represent the data distribution.

For example: we have n observations on two variables X_1 and X_2, where $X1$ is the estimation and X_2 is the ground-truth. $r_{1,2}$ is the Pearson correlation coefficient between X_1 and X_2. The null hypothesis is that there is no significant correlation between X_1 and X_2, i.e. X_1 is an estimation of X_2 that is not better than random.

The p-value of the permutation test is equal to the proportion of sample permutations of X_1 for which the Pearson correlation with X_2 is larger than $r_{1,2}$.

In our example and for a statistical significance level $\alpha = 5\%$: if p-value< 0.05, the null hypothesis is rejected, which means that the correlation between X_1 and X_2 is statistically significant.

Steiger's Z-test: is there a statistically significant difference between two correlated correlation?

The Steiger's Z-test, proposed by Meng et al. [Meng 92], allows to test if the difference between two correlated correlation coefficients is statistically significant, in the case where the coefficient values are not normally distributed.

For example: we have n observations on three variables X_1, X_2 and X_3, where X_1 and X_2 are two estimations and X_3 is the ground-truth. $r_{1,2}$ is the Kendall correlation coefficient between the variables X_1 and X_2, and $r_{1,3}$ is the Kendall correlation coefficient between the variables X_1 and X_3. Then X_3 is the dependent variable and the statistical test considers the Kendall correlation coefficient $r_{2,3}$ between the variables X_2 and X_3. The null hypothesis is that there no significant difference between the two correlation coefficients.

The Steiger's Z-test is the equivalent of the Hotelling's t test [Sheskin 11] for non-normally distributed data: it uses the Fisher's transformation is $z_{i,j} = \frac{1}{2}\ \ln(\frac{1+r_{i,j}}{1-r_{i,j}})$ which converts the correlation coefficients to a normal distribution. The Steiger's Z-test statistic is given by:

$$Z = (z_{1,2} - z_{1,3}) \sqrt{\frac{n-3}{2\ (1-r_{2,3})\ (\frac{1-f\ r_{\text{avg}}}{1-r_{\text{avg}}})}} \tag{4}$$

with $r_{\text{avg}} = \frac{r_{1,2}^2 + r_{1,3}^2}{2}$ and $f = \frac{1-r_{2,3}}{2\ (1-r_{\text{avg}})}$. Under the null hypothesis, Z has a normal distribution and the associated p-value provides the statistical relevance.

In our example and for a statistical significance level $\alpha = 5\%$: if p-value< 0.05, the null hypothesis is rejected, which means that the difference between the two correlation coefficients $r_{1,2}$ and $r_{1,3}$ is statistically significant.

Appendix B: DDW 2010 Clinical Abstract

Based on: [André 10b] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *Endomicroscopic video retrieval approach to support diagnostic differentiation between neoplastic and non-neoplastic colonic polyps.* Gastroenterology (DDW 2010), volume 138, number 5 Suppl, pages S–514, May 2010

Background: Probe-based confocal laser endomicroscopy (pCLE) enables dynamic imaging of the gastrointestinal epithelium In Vivo during ongoing endoscopy and, as of today, relies on the endoscopist for image understanding. The subjective nature of pCLE video semantics suggests the need for a standardized and more automated method for image sequence interpretation.

Aims: To support the diagnosis of a newly acquired pCLE video, we aim at retrieving from a training database videos that have a similar appearance to the video of interest and that have been previously diagnosed by expert physicians with confirmed histology. As a model system, we used distinction of adenomatous and hyperplastic colorectal polyps.

Methods: 68 patients underwent colonoscopy with pCLE for fluorescein-aided imaging of suspicious colonic polyps before their removal. The resulting database is composed of 121 videos (36 non-neoplastic, 85 neoplastic) and 499 edited video sub-sequences (231 non-neoplastic, 268 neoplastic) annotated by clinical experts with a pathological diagnosis. To quantify the relevance of video retrieval, we performed an unbiased classification with leave-one-patient-out cross-validation, based on the voting of the k most similar videos. The Bag-of-Visual-Words method from computer vision extracts local continuous image features and clusters them into a finite number of visual words to build an efficient image signature. In order to retrieve videos and not only isolated images, we revisited this method and analyzed the impact of including spatial overlap between time-related images. We first used the results of a video-mosaicing technique to weight the contribution of each local image region to its visual word. Then, we computed the video signatures with a histogram summation technique, which reduces both retrieval runtime and training memory.

Results: Video classification results show that our method achieves, when using the votes of the $k = 9$ most similar videos, a sensitivity of 97.7% and a specificity of 86.1% for a resulting accuracy of 94.2%. When compared to using the still images independently, using video data improves the results in a statistically significant

manner (McNemar's test: p-value= 0.021 when using the votes of the $k = 3$ most similar videos). Moreover, fewer similar videos are necessary to classify the query at a given accuracy, which is clinically relevant for the physician.

Conclusion: Our method using the results of video-mosaicing for content-based video retrieval appears to be highly accurate for pCLE videos. It may provide the endoscopist with diagnostic decision support and avoid unnecessary polypectomy of non-neoplastic lesions.

Appendix C: DDW 2011 Clinical Abstract

Based on: [André 11b] B. André, T. Vercauteren, A. M. Buchner, M. W. Shahid, M. B. Wallace and N. Ayache. *Toward a structured training system for probe-based confocal laser endomicroscopy (pCLE) on Barrett's esophagus: a video retrieval approach to estimate diagnosis difficulty.* Gastrointestinal Endosccopy (DDW 2011), volume 73, number 4 Suppl, page AB398, 2011. Selected Video Abstract available at `http://www.youtube.com/watch?v=RVy-0Bxx9EQ`

Background: pCLE (Cellvizio, Mauna Kea Technologies) allows the endoscopist to image the epithelial surface *in vivo*, at microscopic level and in real-time (12 frames per second) during an ongoing endoscopy. Early diagnosis of epithelial cancers with pCLE may be perceived as a challenging task for many new endoscopists. There is a crucial need to provide objective methods to diagnose neoplasia, estimate confidence levels, and to shorten the learning curve.

Aims: Our long-term objective is to develop a modular training system for pCLE diagnosis, by adapting the difficulty level according to the endoscopist's expertise. This study aims at providing an automated estimation of the diagnosis difficulty. As the understanding of pCLE video diagnosis is driven by perceived visual similarity, we propose a content-based video retrieval approach toward this goal.

Methods: Our database contains annotated pCLE videos of BE that were provided by the multicentric study NCT00795184. It includes 76 patients and 123 videos (62 benign, 61 neoplastic) split into 862 stable video subsequences. 20 of these videos (9 benign, 11 neoplastic) were graded offline by 21 endoscopists, including 9 pCLE experts and 12 non-experts, who individually established a blinded pCLE diagnosis for each lesion. A single expert GI pathologist reviewed all the biopsies acquired on the imaging spots and provided a reference diagnosis. The percentage of false pCLE diagnosis established on a video among the endoscopists is our "ground truth" for the diagnosis difficulty of the video. We first applied to the video database a video retrieval method that we developed especially for this task. We then used the retrieval results to extract a relevant difficulty criterion that measures contextual discrepancies between the video query and its most visually similar videos.

Results: Our video retrieval method, objectively evaluated using k-nearest neighbor classification, outperforms several state-of-the-art methods on the BE

database (acc. 85.4%, sens. 90.2%, spec. 80.7%). Our estimated diagnosis difficulty has a correlation of 0.78 (p-value < 0.0002) with the ground truth difficulty measured for all the endoscopists, 0.63 (p-value < 0.003) with those measured for the experts only, and 0.80 (p-value < 0.0001) with those measured for the non-experts only.

Conclusion: Our experiments demonstrate that there is a noticeable relationship between our retrieval-based difficulty estimation and the difficulty experienced by the endoscopists. The complete video database with estimated difficulty could thus be used to identify lesions for which an optical diagnosis will be difficult, and to develop a training simulator that features difficulty level selection. Finally, a clinical validation will be required to assess whether such a structured training system will eventually help shorten the pCLE learning curve.

Bibliography

[Agarwal 08] A. Agarwal and B. Triggs. *Multilevel image coding with hyperfeatures.* International Journal of Computer Vision, volume 78, number 1, pages 15–27, 2008. Cited on page(s) 55.

[Akaike 74] H Akaike. *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, volume 19, number 6, pages 716–723, December 1974. Cited on page(s) 84.

[Akgül 11] C. B. Akgül, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan and B. Acar. *Content-based image retrieval in radiology: Current status and future directions.* Journal of Digital Imaging, volume 24, number 2, pages 208–222, 2011. Cited on page(s) 89.

[Allain 10] B. Allain, M. Hu, L. B. Lovat, R. J. Cook, T. Vercauteren, S. Ourselin and D. J. Hawkes. *A system for biopsy site retargeting with uncertainty in gastroenterology and oropharyngeal examinations.* In Proceedings of the 13th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'10), pages 514–521, 2010. Cited on page(s) 119.

[André 09a] B. André, T. Vercauteren, A. Perchant, M. B. Wallace, A. M. Buchner and N. Ayache. *Endomicroscopic image retrieval and classification using invariant visual features.* In Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'09), pages 346–349, 2009. Cited on page(s) 9, 11, 22, 116.

[André 09b] B. André, T. Vercauteren, A. Perchant, M. B. Wallace, A. M. Buchner and N. Ayache. *Introducing space and time in local feature-based endomicroscopic image retrieval.* In Proceedings of the MICCAI 2009 Workshop - Medical Content-based Retrieval for Clinical Decision (MCBR-CDS'09), pages 18–30, 2009. Cited on page(s) 9, 11, 31, 116.

[André 10a] B. André, T. Vercauteren, A. M. Buchner, M. W. Shahid, M. B. Wallace and N. Ayache. *An image retrieval approach to setup difficulty levels in training systems for endomicroscopy*

diagnosis. In Proceedings of the 13th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'10), pages 480–487, 2010. Cited on page(s) 7, 9, 11, 75, 116.

[André 10b] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *Endomicroscopic video retrieval approach to support diagnostic differentiation between neoplastic and non-neoplastic colonic polyps*. Gastroenterology (DDW 2010), volume 138, number 5 Suppl, pages S–514, May 2010. Cited on page(s) 7, 10, 57, 116, 124.

[André 10c] B. André, T. Vercauteren, M. B. Wallace, A. M. Buchner and N. Ayache. *Endomicroscopic video retrieval using mosaicing and visual words*. In Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'10), pages 1419–1422, 2010. Cited on page(s) 9, 37, 116.

[André 11a] B. André, T. Vercauteren, A. M. Buchner, M. Krishna, N. Ayache and M. B. Wallace. *Video retrieval software for automated classification of probe-based confocal laser endomicroscopy on colorectal polyps*. 2011. Article in submission. Cited on page(s) 7, 9, 57, 116.

[André 11b] B. André, T. Vercauteren, A. M. Buchner, M. W. Shahid, M. B. Wallace and N. Ayache. *Toward a structured training system for probe-based confocal laser endomicroscopy (pCLE) on Barrett's esophagus: a video retrieval approach to estimate diagnosis difficulty*. Gastrointestinal Endosccopy (DDW 2011), volume 73, number 4 Suppl, page AB398, 2011. Selected Video Abstract available at `http://www.youtube.com/watch?v=RVy-0Bxx9EQ`. Cited on page(s) 7, 10, 75, 126.

[André 11c] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *Learning semantic and visual similarity for endomicroscopy video retrieval*. 2011. Article in submission. Cited on page(s) 7, 9, 87, 116, 117.

[André 11d] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *Retrieval evaluation and distance learning from perceived similarity between endomicroscopy videos*. In Proceedings of the 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'11), pages 289–296, 2011. Cited on page(s) 9, 87, 89, 93, 116, 117.

[André 11e] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace and N. Ayache. *A smart atlas for endomicroscopy using automated video retrieval.* Medical Image Analysis, volume 15, number 4, pages 460–476, August 2011. Cited on page(s) 7, 9, 11, 116.

[Barillot 04] C. Barillot, R. Valabregue, J-P. Matsumoto, F. Aubry, H. Benali, Y. Cointepas, O. Dameron, M. Dojat, E. Duchesnay, B. Gibaud, S. Kinkingnéhun, D. Papadopoulos, M. Pellegrini-Issac and E. Simon. *Neurobase: Management of distributed and heterogeneous information sources in neuroimaging.* In M. Dojat and B. Gibaud, editors, DiDaMIC Workshop, MICCAI 2004 Conference, pages 85–94, 2004. Cited on page(s) 6.

[Barnett 91] V. Barnett. Sample survey principles and methods. Hodder Arnold, 1991. Cited on page(s) 104.

[Bay 06] H. Bay, T. Tuytelaars and L. J. Van Gool. *SURF: Speeded Up Robust Features.* In Proceedings of the 9th European Conference on Computer Vision (ECCV'06), pages 404–417, 2006. Cited on page(s) 24.

[Becker 07] V. Becker, T. Vercauteren, C. H. von Weyern, C. Prinz, R. M. Schmid and A. Meining. *High resolution miniprobe-based confocal microscopy in combination with video-mosaicing.* Gastrointestinal Endoscopy, volume 66, number 5, pages 1001–1007, November 2007. Cited on page(s) 27, 66.

[Boiman 08] O. Boiman, E. Shechtman and M. Irani. *In defense of nearest-neighbor based image classification.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), pages 1–8, 2008. Cited on page(s) 20, 49, 81.

[Boureau 10] Y-L. Boureau, F. Bach, Y. LeCun and J. Ponce. *Learning mid-level features for recognition.* In Proceedings of the International Conference on Computer Vision and Pattern Recognition, pages 2559–2566, 2010. Cited on page(s) 6.

[Brydges 09] R. N. Brydges. A critical reappraisal of self-learning in health professions education: Directed self-guided learning using simulation modalities. University of Toronto, 2009. Cited on page(s) 77.

[Buchner 08] A. M. Buchner, M. S. Ghabril, M. Krishna, H. C. Wolfsen and M. B. Wallace. *High-resolution confocal endomi-*

croscopy probe system for in vivo diagnosis of colorectal neoplasia. Gastroenterology, volume 135, number 1, page 295, July 2008. Cited on page(s) 18.

[Buchner 09a] A. M. Buchner, V. Gomez, K. R. Gill, M. Ghabril, D. Scimeca, M. W. Shahid, S. R. Achem, M. F. Picco, D. Riegert-Johnson, M. Raimondo, H. C. Wolfsen, T. A. Woodward, M. K. Hasan and M. B. Wallace. *The learning curve for in vivo probe based Confocal Laser Endomicroscopy (pCLE) for prediction of colorectal neoplasia.* Gastrointestinal Endoscopy, volume 69, number 5, pages AB364–AB365, April 2009. Cited on page(s) 55, 73, 79.

[Buchner 09b] A. M. Buchner, M. W. Shahid, M. G. Heckman, M. Krishna, M. Ghabril, M. Hasan, J. E. Crook, V. Gomez, M. Raimondo, T. Woodward, H.C. Wolfsen and M. B. Wallace. *Comparison of probe-based confocal laser endomicroscopy with virtual chromoendoscopy for classification of colon polyps.* Gastroenterology, volume 138, number 3, pages 834–842, November 2009. Cited on page(s) 18.

[Buchner 10] A. M. Buchner, M. W. Shahid, M. G. Heckman, M. Krishna, M. Ghabril, M. Hasan, J. E. Crook, V. Gomez, M. Raimondo, T. Woodward, H. C. Wolfsen and M. B. Wallace. *Comparison of probe-based confocal laser endomicroscopy with virtual chromoendoscopy for classification of colon polyps.* Gastroenterology, volume 138, number 3, pages 834–42, 2010. Cited on page(s) 60.

[Caicedo 10] J. C. Caicedo, J. G. Moreno, E. A. Niño and F. A. González. *Combining visual features and text data for medical image retrieval using latent semantic kernels.* In Proceedings of Multimedia Information Retrieval, pages 359–366, 2010. Cited on page(s) 96.

[Chehade 09] N. H. Chehade, J-G. Boureau, C. Vidal and J. Zerubia. *Multiclass SVM for forestry classification.* In Proceedings of the International Conference on Image Processing, pages 1673–1676, 2009. Cited on page(s) 6.

[Dabizzi 11] E. Dabizzi, M. W. Shahid, B. Qumseya, M. Othman and M. B. Wallace. *Comparison between video and mosaics viewing mode of confocal laser endomicroscopy (pCLE) in patients with Barrett's esophagus.* Gastrointestinal Endoscopy (DDW 2011), volume 73, number 4 Suppl, page AB282, 2011. Cited on page(s) 92.

[De Palma 10] G. D. De Palma, S. Staibano, S. Siciliano, M. Persico, S. Masone, F. Maione, M. Siano, M. Mascolo, D. Esposito, F. Salvatori and G. Persico. *In vivo characterisation of superficial colorectal neoplastic lesions with high-resolution probe-based confocal laser endomicroscopy in combination with video-mosaicing: a feasibility study to enhance routine endoscopy.* Digestive and Liver Disease, volume 42, number 11, pages 791–7, 2010. Cited on page(s) 27, 92.

[de Visser 10] H. de Visser, J. Passenger, D. Conlan, C. Russ, D. Hellier, M. Cheng, O. Acosta, S. Ourselin and O. Salvado. *Developing a next generation colonoscopy simulator.* International Journal of Image and Graphics, volume 10, number 2, pages 203–217, 2010. Cited on page(s) 77.

[Descombes 99] X. Descombes, R. Morris, J. Zerubia and M. Berthod. *Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood.* IEEE Transactions on Image Processing, volume 8, number 7, pages 954–963, July 1999. Cited on page(s) 55.

[Désir 10] C. Désir, C. Petitjean, L. Heutte and L. Thiberville. *Using a priori knowledge to classify in vivo images of the lung.* In Intelligent Computing in Image Processing, pages 207–212, 2010. Cited on page(s) 18.

[DONT BIOPCE 10] DONT BIOPCE. *Detection Of Neoplastic Tissue in Barrett's esophagus with In vivO Probe-based Confocal Endomicroscopy (DONT BIOPCE)*, June 2010. `http://clinicaltrials.gov/ct2/show/NCT00795184`. Cited on page(s) 79.

[Doyle 06] S. Doyle, A. Madabhushi, M. D. Feldman and J. E. Tomaszeweski. *A boosting cascade for automated detection of prostate cancer from digitized histology.* In Proceedings of the 9th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'06), pages 504–511, 2006. Cited on page(s) 15.

[Dundar 04] M. Dundar, G. Fung, L. Bogoni, M. Macari, A. Megibow and R. B. Rao. *A methodology for training and validating a CAD system and potential pitfalls.* In Computer Assisted Radiology and Surgery, pages 1010–1014, 2004. Cited on page(s) 19, 67.

[El-Naqa 04] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa and M. N. Wernick. *A similarity learning approach to content-*

based image retrieval: application to digital mammography. IEEE Transactions on Medical Imaging, volume 23, number 10, pages 1233–1244, 2004. Cited on page(s) 99.

[Gomez 10] V. Gomez, A. M. Buchner, E. Dekker, F. J. van den Broek, A. Meining, M. W. Shahid, M. Ghabril, P. Fockens, M. G. Heckman and M. B. Wallace. *Interobserver agreement and accuracy among international experts with probe-based confocal laser endomicroscopy in predicting colorectal neoplasia.* Endoscopy, volume 42, number 4, pages 286–291, 2010. Cited on page(s) 55, 73, 79.

[Greenspan 09] H. Greenspan. *Revisiting the feature and content gap for landmark-based and image-to-image retrieval in medical CBIR.* International Journal of Healthcare Information Systems and Informatics, volume 4, number 1, pages 68–87, 2009. Cited on page(s) 6.

[Gurcan 09] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. Rajpoot and B. Yener. *Histopathological image analysis: A review.* IEEE Reviews in Biomedical Engineering, volume 2, pages 147–171, 2009. Cited on page(s) 15.

[Häfner 09] M. Häfner, A. Gangl, R. Kwitt, A. Uhl, A. Vécsei and F. Wrba. *Improving pit-pattern classification of endoscopy images by a combination of experts.* In Proceedings of the 12th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'09), pages 247–254, 2009. Cited on page(s) 18.

[Haralick 79] R. M. Haralick. *Statistical and structural approaches to texture.* In Proceedings of the IEEE, volume 67, pages 786–804, 1979. Cited on page(s) 6, 18, 96.

[Hauff 08] C. Hauff, D. Hiemstra and F. de Jong. *A survey of pre-retrieval query performance predictors.* In Proceedings of the International Conference on Information and Knowledge Management, pages 1419–1420, 2008. Cited on page(s) 78.

[Hawk 05] E. T. Hawk and B. Levin. *Colorectal cancer prevention.* J Clin Oncol, volume 23, number 2, pages 378–91, 2005. Cited on page(s) 59.

[Hoi 08] S. C. H. Hoi, R. Jin, J. Zhu and M. R. Lyu. *Semi-supervised SVM batch mode active learning for image retrieval.* In Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition (CVPR'08), pages 24–26, 2008. Cited on page(s) 84.

[Hurlstone 08] D.P. Hurlstone. *Surface analysis with magnifying chromoendoscopy in the colon.* In Atlas of Endomicroscopy, chapitre 2, pages 7–15. Springer Berlin Heidelberg, 2008. Cited on page(s) 62, 63.

[Iglesias 11] J. E. Iglesias, E. Konukoglu, A. Montillo, Z. Tu and A. Criminisi. *Combining generative and discriminative models for semantic segmentation of CT scans via active learning.* In IPMI, pages 25–36, 2011. Cited on page(s) 119.

[Jegou 08] H. Jegou, M. Douze and C. Schmid. *Hamming embedding and weak geometric consistency for large scale image search.* In Proceedings of the 10th European Conference on Computer Vision (ECCV'08), volume I, pages 304–317, October 2008. Cited on page(s) 30.

[Jones 96] B. Jones, P. Jarvis, J. A. Lewis and A. F. Ebbutt. *Trials to assess equivalence: the importance of rigorous methods.* British Medical Journal, volume 313, number 7048, pages 36–9, 1996. Cited on page(s) 67, 122.

[Khalid 09] O. Khalid, S. Radaideh, O. W. Cummings, M. J. O' Brien, J. R. Goldblum and D. K. Rex. *Reinterpretation of histology of proximal colon polyps called hyperplastic in 2001.* World Journal of Gastroenterology, volume 15, number 30, pages 3767–70, 2009. Cited on page(s) 73, 89.

[Kiesslich 04] R. Kiesslich, J. Burg, M. Vieth, J. Gnaendiger, M. Enders, P. Delaney, A. Polglase, W. McLaren, D. Janell, S. Thomas, B. Nafe, P. R. Galle and M. F. Neurath. *Confocal laser endoscopy for diagnosing intraepithelial neoplasias and colorectal cancer in vivo.* Gastroenterology, volume 127, number 3, pages 706–13, 2004. Cited on page(s) 62, 95.

[Kong 09] J. Kong, O. Sertel, H. Shimada, K. L. Boyer, J. H. Saltz and M. N. Gurcan. *Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation.* Pattern Recognition, volume 42, number 6, pages 1080–1092, 2009. Cited on page(s) 15.

[Kudo 96] S. Kudo, S. Tamura, T. Nakajima, H. Yamano, H. Kusaka and H. Watanabe. *Diagnosis of colorectal tumorous lesions by magnifying endoscopy.* Gastrointestinal Endoscopy, volume 44, number 1, pages 8–14, 1996. Cited on page(s) 61.

[Kwitt 11] R. Kwitt, N. Rasiwasia, N. Vasconcelos, A. Uhl, M. Häfner and F. Wrba. *Learning pit pattern concepts for gastroenterological training.* In Proceedings of the 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'11), pages 273–280, 2011. Cited on page(s) 96.

[Laptev 08] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld. *Learning realistic human actions from movies.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), pages 1–8, 2008. Cited on page(s) 15.

[Lazebnik 06] S. Lazebnik, C. Schmid and P. Ponce. *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories.* In Proceedings of the International Conference on Computer Vision and Pattern Recognition, pages 2169–2178, 2006. Cited on page(s) 6, 30.

[Le Goualher 04] G. Le Goualher, A. Perchant, M. Genet, C. Cavé, B. Viellerobe, F. Berier, B. Abrat and N. Ayache. *Towards optical biopsies with an integrated fibered confocal fluorescence microscope.* In Proceedings of the 7th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'04), pages 761–768, 2004. Cited on page(s) 17.

[Leung 01] T. Leung and J. Malik. *Representing and recognizing the visual appearance of materials using three-dimensional textons.* International Journal of Computer Vision, volume 43, pages 29–44, June 2001. Cited on page(s) 16, 18, 81, 96.

[Long 09] L. R. Long, S. Antani, T. M. Deserno and G. R. Thoma. *Content-based image retrieval in medicine: Retrospective assessment, state of the art, and future directions.* International journal of healthcare information systems and informatics, volume 4, number 1, pages 1–16, 2009. Cited on page(s) 15.

[Lowe 04] D. G. Lowe. *Distinctive image features from scale-invariant keypoints.* International Journal of Computer Vision, volume 60, pages 91–110, November 2004. Cited on page(s) 6, 22.

[Ma 10] H. Ma, J. Zhu, M. R. Lyu and I. King. *Bridging the semantic gap between image contents and tags.* IEEE Transactions on Multimedia, volume 12, pages 462–473, 2010. Cited on page(s) 96.

[Matas 04] J. Matas, O. Chum, M. Urban and T. Pajdla. *Robust wide baseline stereo from maximally stable extremal regions.* Image and Vision Computing, volume 22, number 10, pages 761–767, 2004. Cited on page(s) 23.

[Meining 07] A. Meining, D. Saur, M. Bajbouj, V. Becker, E. Peltier, H. Höfler, C. H. von Weyhern, R. M. Schmid and C. Prinz. *In vivo histopathology for detection of gastrointestinal neoplasia with a portable, confocal miniprobe: an examiner blinded analysis.* Clinical Gastroenterology and Hepatology, volume 5, number 11, pages 1261–7, 2007. Cited on page(s) 60.

[Meng 92] X-L. Meng, R. Rosenthal and D. B. Rubin. *Comparing correlated correlation coefficients.* Psychological Bulletin, volume 111, number 1, pages 172–175, 1992. Cited on page(s) 104, 123.

[Mikolajczyk 05] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. J. Van Gool. *A comparison of affine region detectors.* International Journal of Computer Vision, volume 65, pages 43–72, November 2005. Cited on page(s) 23.

[Modat 10] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox and S. Ourselin. *Fast free-form deformation using graphics processing units.* Computer Methods and Programs in Biomedicine, volume 98, number 3, pages 278–284, 2010. Cited on page(s) 37.

[Muja 09] M. Muja and D. G. Lowe. *Fast approximate nearest neighbors with automatic algorithm configuration.* In VISAPP, pages 331–340, 2009. Cited on page(s) 22.

[Müller 04] H. Müller, N. Michoux, D. Bandon and D. Geissbühler. *A review of content-based image retrieval systems in medical applications - clinical benefits and future directions.* International Journal of Medical Informatics, volume 73, number 1, pages 1–23, 2004. Cited on page(s) 15.

[Müller 08] H. Müller, J. Kalpathy-Cramer, C. E. Kahn, W. Hatt, S. Bedrick and W. R. Hersh. *Overview of the ImageCLEFmed 2008 medical image retrieval task.* In CLEF, pages 512–522, 2008. Cited on page(s) 6, 15.

[Nister 06] D. Nister and H. Stewenius. *Scalable recognition with a vocabulary tree.* In Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (CVPR'06), pages 2161–2168, 2006. Cited on page(s) 22.

[Norfleet 88] R. G. Norfleet, M. E. Ryan and J. B. Wyman. *Adenomatous and hyperplastic polyps cannot be reliably distinguished by their appearance through the fiberoptic sigmoidoscope*. Digestive Diseases and Sciences, volume 33, number 9, pages 1175–7, 1988. Cited on page(s) 13, 60.

[Oliva 01] A. Oliva and A. Torralba. *Modeling the shape of the scene: A holistic representation of the spatial envelope*. International Journal of Computer Vision, volume 42, number 3, pages 145–175, 2001. Cited on page(s) 6.

[ParisWorkshop 03] ParisWorkshop. *The Paris endoscopic classification of superficial neoplastic lesions: Esophagus, stomach, and colon: November 30 to December 1, 2002*. Gastrointestinal Endoscopy, volume 58, number 6 Suppl, pages S3–S43, 2003. Cited on page(s) 61.

[Pele 09] O. Pele and M. Werman. *Fast and robust earth mover's distances*. In Proceedings of the 19th International Conference on Computer Vision (ICCV'09), 2009. Cited on page(s) 22.

[Pernod 11] E. Pernod, M. Sermesant, E. Konukoglu, J. Relan, H. Delingette and N. Ayache. *A multi-front eikonal model of cardiac electrophysiology for interactive simulation of radiofrequency ablation*. Computers and Graphics, volume 35, pages 431–440, 2011. Cited on page(s) 77.

[Perronnin 07] F. Perronnin and C. Dance. *Fisher kernels on visual vocabularies for image categorization*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), pages 1–8, 2007. Cited on page(s) 29.

[PERSEE 10] PERSEE. *Persee project*, 2010. `https://persee.maunakeatech.com/Public/press`. Cited on page(s) 120.

[Petrou 06] M. Petrou, R. Piroddi and A. Talepbour. *Texture recognition from sparsely and irregularly sampled data*. Computer vision and image understanding, volume 102, number 1, pages 95–104, 2006. Cited on page(s) 18.

[Philbin 10] J. Philbin, M. Isard, J. Sivic and A. Zisserman. *Descriptor learning for efficient retrieval*. In Proceedings of the 11th European Conference on Computer Vision (ECCV'10), pages 677–691, 2010. Cited on page(s) 99.

[Poblete 10] B. Poblete, B. Bustos, M. Mendoza and J. M. Barrios. *Visual-semantic graphs: using queries to reduce the semantic gap in web image retrieval*. In Proceedings of the ACM International Conference on Information and Knowledge Management, pages 1553–1556, 2010. Cited on page(s) 96.

[Pohl 08] H. Pohl, T. Rosch, M. Vieth, M. Koch, V. Becker, M. Anders, A. Khalifa and A. Meining. *Miniprobe confocal laser microscopy for the detection of invisible neoplasia in patients with Barrett's esophagus*. Gut, volume 57, number 12, pages 1648–1653, 2008. Cited on page(s) 54.

[Rasiwasia 07] N. Rasiwasia, P. J. Moreno and N. Vasconcelos. *Bridging the gap: Query by semantic example*. IEEE Transactions on Multimedia, volume 9, number 5, pages 923–938, 2007. Cited on page(s) 96.

[Rasiwasia 10] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy and N. Vasconcelos. *A new approach to cross-modal multimedia retrieval*. In Proceedings of the ACM International Conference on Multimedia, pages 251–260, 2010. Cited on page(s) 96.

[Rastogi 09] A. Rastogi, J. Keighley, V. Singh, P. Callahan, A. Bansal, S. Wani and P. Sharma. *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*. The American Journal of Gastroenterology, volume 104, number 10, pages 2422–30, 2009. Cited on page(s) 13, 60.

[Rex 11] D. K. Rex, C. Kahi, M. O' Brien, T. R. Levin, H. Pohl, A. Rastogi, L. Burgart, T. Imperiale, U. Ladabaum, J. Cohen and D. A. Lieberman. *The American Society for Gastrointestinal Endoscopy PIVI (Preservation and Incorporation of Valuable Endoscopic Innovations) on real-time endoscopic assessment of the histology of diminutive colorectal polyps*. Gastrointestinal Endoscopy, volume 73, number 3, pages 419–422, March 2011. Cited on page(s) 117.

[Rhienmora 11] P. Rhienmora, P. Haddawy, S. Suebnukarn and M. N. Dailey. *Intelligent dental training simulator with objective skill assessment and feedback*. Artificial Intelligence in Medicine, volume 52, number 2, pages 115–121, 2011. Cited on page(s) 76.

[Rubio 06] C. A. Rubio, G. Nesi, L. Messerini, G. C. Zampi, K. Mandai, M. Itabashi and K. Takubo. *The Vienna classification applied to colorectal adenomas*. Journal of Gastroenterology and Hepatology, volume 21, number 11, pages 1697–703, 2006. Cited on page(s) 64.

[Rubner 00] Y. Rubner, C. Tomasi and L. J. Guibas. *The Earth Mover's Distance as a metric for image retrieval*. International Journal of Computer Vision, volume 40, number 2, pages 99–121, November 2000. Cited on page(s) 22.

[Salton 88] G. Salton and C. Buckley. *Term-weighting approaches in automatic text retrieval*. In Information Processing and Management, pages 513–523, 1988. Cited on page(s) 78.

[Savarese 06] S Savarese, J. M. Winn and A Criminisi. *Discriminative object class models of appearance and shape by correlatons*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06), pages 2033–2040, 2006. Cited on page(s) 33.

[Schlemper 00] R. J. Schlemper, R. H. Riddell, Y. Kato, F. Borchard, H. S. Cooper, S. M. Dawsey, M. F. Dixon, C. M. Fenoglio-Preiser, J. F. Fléjou, K. Geboes, T. Hattori, T. Hirota, M. Itabashi, M. Iwafuchi, A. Iwashita, Y.I. Kim, T. Kirchner, M. Klimpfinger, M. Koike, G.Y. Lauwers, K.J. Lewin, G. Oberhuber, F. Offner, A.B. Price, C.A. Rubio, M. Shimizu, T. Shimoda, P. Sipponen, E. Solcia, M. Stolte, H. Watanabe and H. Yamabe. *The Vienna classification of gastrointestinal epithelial neoplasia*. Gut, volume 47, number 2, pages 251–5, 2000. Cited on page(s) 64.

[Scholer 09] F. Scholer and S. Garcia. *A case for improved evaluation of query difficulty prediction*. In Proceedings of the ACM Special Interest Group on Information Retrieval, pages 640–641, 2009. Cited on page(s) 78.

[Schwaninger 07] A. Schwaninger, S. Michel and A. Bolfing. *A statistical approach for image difficulty estimation in X-ray screening using image measurements*. In Proceedings of the Symposium on Applied Perception in Graphics and Visualisation, pages 123–130, 2007. Cited on page(s) 78.

[Schwarz 78] G. Schwarz. *Estimating the dimension of a model*. The Annals of Statistics, volume 6, pages 461–464, 1978. Cited on page(s) 84.

[Scovanner 07] P. Scovanner, S. Ali and M. Shah. *A 3-dimensional SIFT descriptor and its application to action recognition.* In Proceedings of the ACM International Conference on Multimedia, pages 357–360, 2007. Cited on page(s) 55.

[Sharma 11] P. Sharma, A. R. Meining, E. Coron, C. J. Lightdale, H. C. Wolfsen, A. Bansal, M. Bajbouj, J. P. Galmiche, J. A. Abrams, A. Rastogi, N. Gupta, J. E. Michalek, G. Y. Lauwers and M. B. Wallace. *Real-time increased detection of neoplastic tissue in Barrett's esophagus with probe-based confocal laser endomicroscopy: final results of an international multicenter, prospective, randomized, controlled trial.* Gastrointest Endoscopy, 2011. Cited on page(s) 54.

[Sheskin 11] D. J. Sheskin. Handbook of parametric and nonparametric statistical procedures. Chapman & Hall/CRC, 5th Revised edition, 2011. Cited on page(s) 19, 67, 121, 123.

[Shotton 06] J. Shotton, J. M. Winn, C. Rother and A. Criminisi. *TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation.* In Proceedings of the 9th European Conference on Computer Vision (ECCV'06), pages 1–15, 2006. Cited on page(s) 23.

[Simonyan 11] K. Simonyan, A. Zisserman and A. Criminisi. *Immediate structured visual search for medical images.* In Proceedings of the 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'11), pages 281–288, 2011. Cited on page(s) 6.

[Sivic 06] J. Sivic and A. Zisserman. *Video Google: Efficient visual search of videos.* In Toward Category-Level Object Recognition, pages 127–144, 2006. Cited on page(s) 6, 15, 22.

[Sivic 09] J Sivic and A Zisserman. *Efficient visual search of videos cast as text retrieval.* IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 31, number 4, pages 591–606, 2009. Cited on page(s) 29.

[Smeulders 00] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain. *Content-based image retrieval at the end of the early years.* IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 22, number 12, pages 1349–1380, 2000. Cited on page(s) 6, 15, 18, 89.

[Srivastava 08] S. Srivastava, J. J. Rodriguez, A. R. Rouse, M. A. Brewer and A. F. Gmitro. *Computer-aided identification of ovarian can-*

cer in confocal microendoscope images. Journal of Biomedical Optics, volume 13, number 2, page 024021, March/April 2008. Cited on page(s) 18.

[Thiberville 07] L. Thiberville, S. Moreno-Swirc, T. Vercauteren, E. Peltier, C. Cavé and G. Bourg Heckly. *In vivo imaging of the bronchial wall microstructure using fibered confocal fluorescence microscopy*. American Journal of Respiratory and Critical Care Medicine, volume 175, number 1, pages 22–31, January 2007. Cited on page(s) 27.

[Turpin 06] A. Turpin and F. Scholer. *User performance versus precision measures for simple search tasks*. In Proceedings of the Special Interest Group on Information Retrieval, pages 11–18, 2006. Cited on page(s) 78.

[Tuytelaars 00] T. Tuytelaars and L. J. Van Gool. *Wide baseline stereo matching based on local, affinely invariant regions*. In British Machine Vision Conference, 2000. Cited on page(s) 23.

[Tuytelaars 08] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: A survey. Now Publishers Inc., 2008. Cited on page(s) 23.

[Vercauteren 06] T. Vercauteren, A. Perchant, G. Malandain, X. Pennec and N. Ayache. *Robust mosaicing with correction of motion distortions and tissue deformation for in vivo fibered microscopy*. Medical Image Analysis, volume 10, number 5, pages 673–692, October 2006. Cited on page(s) 14, 16, 27, 66, 92.

[Vercauteren 08] T. Vercauteren, A. Meining, F. Lacombe and A Perchant. *Real time autonomous video image registration for endomicroscopy: Fighting the compromises*. In Proceedings of the SPIE BIOS - Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XV, volume 6861, page 68610C, January 2008. Cited on page(s) 16, 37.

[VSS 11] VSS. *Visual similarity scoring (VSS)*, 2011. `http://smartatlas.maunakeatech.com`, login: MICCAI-User, password: MICCAI2011. Cited on page(s) 93.

[Wallace 09] M. B. Wallace and P. Fockens. *Probe-based confocal laser endomicroscopy*. Gastroenterology, volume 136, number 5, pages 1509–1513, May 2009. Cited on page(s) 13, 60.

[Wang 07] K. K. Wang and M. Camilleri. *Endoscopic confocal microscopy: imaging to facilitate the dawn of endoluminal*

surgery. Clinical Gastroenterology and Hepatology, volume 5, number 11, pages 1259–1260, 2007. Cited on page(s) 3.

[Wang 09] H. Wang, M. M. Ullah, A. Kläser, I. Laptev and C. Schmid. *Evaluation of local spatio-temporal features for action recognition*. In British Machine Vision Conference, page 127, September 2009. Cited on page(s) 55.

[Weinberger 09] K. Q. Weinberger and L. K. Saul. *Distance metric learning for large margin nearest neighbor classification*. Journal of Machine Learning Research, volume 10, pages 207–244, 2009. Cited on page(s) 99.

[Winawer 06] S. J. Winawer, A. G. Zauber, R. H. Fletcher, J. S. Stillman, M. J. O' Brien, B. Levin, R. A. Smith, D. A. Lieberman, R. W. Burt, T. R. Levin, J. H. Bond, D. Brooks, T. Byers, N. Hyman, L. Kirk, A. Thorson, C. Simmang, D. Johnson and D. K. Rex. *Guidelines for colonoscopy surveillance after polypectomy: a consensus update by the us multi-society task force on colorectal cancer and the american cancer society*. Gastroenterology, volume 130, number 6, pages 1872–85, 2006. Cited on page(s) 60.

[Winn 05] J. M. Winn, A. Criminisi and T. P. Minka. *Object categorization by learned universal visual dictionary*. In Proceedings of the 15th International Conference on Computer Vision (ICCV'05), pages 1800–1807, 2005. Cited on page(s) 29.

[Yang 10] L. Yang, R. Jin, L. Mummert, R. Sukthankar, A. Goode, B. Zheng, S. C. H. Hoi and M. Satyanarayanan. *A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval*. IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 32, pages 30–44, 2010. Cited on page(s) 99.

[Zhang 07] J. Zhang, S. Lazebnik and C. Schmid. *Local features and kernels for classification of texture and object categories: a comprehensive study*. International Journal of Computer Vision, volume 73, pages 213–238, June 2007. Cited on page(s) 6, 15, 18, 22, 24, 65, 80, 96.

[Zhang 09] S. Zhang, Q. Tian, G. Hua, Q. Huang and S. Li. *Descriptive visual words and visual phrases for image applications*. In IEEE Multimedia, pages 75–84, 2009. Cited on page(s) 55.

[Zhao 08] Y. Zhao, F. Scholer and Y. Tsegay. *Effective pre-retrieval query performance prediction using similarity and variability*

evidence. In Proceedings of the European Conference on Information Retrieval, pages 52–64, 2008. Cited on page(s) 78.

Smart Atlas for Endomicroscopy Diagnosis Support: A Clinical Application of Content-Based Image Retrieval

Abstract: Probe-based Confocal Laser Endomicroscopy (pCLE) enables *in vivo* microscopic imaging of the epithelium during ongoing endoscopy, *in situ* and at real-time frame rate. Thanks to this novel imaging system, the endoscopists have the opportunity to perform non-invasive "optical biopsies". Traditional biopsies result in histological images that are usually diagnosed *ex vivo* by pathologists. The *in vivo* diagnosis of pCLE images is therefore a critical challenge for the endoscopists who typically have only little pathology expertise. The main goal of this thesis is to assist the endoscopists in the *in vivo* interpretation of pCLE image sequences.

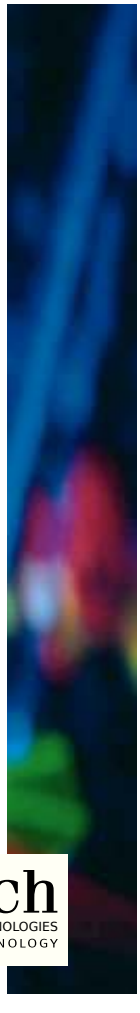
When establishing a diagnosis, physicians typically rely on similarity-based reasoning. To mimic this process, we explore content-based image retrieval (CBIR) approaches for diagnosis support. Our primary objective is to develop a system which automatically extracts several videos that are visually *similar* to the pCLE video of interest, but that are annotated with metadata such as textual diagnosis. Such a retrieval system should help the endoscopist in making an informed decision and therefore a more accurate pCLE diagnosis.

For this purpose, we investigate the Bag-of-Visual-Words (BoW) method from computer vision. Analyzing the image properties of pCLE data leads us to adjust the standard BoW method. Not only single pCLE images, but full pCLE videos are retrieved by representing videos as sets of mosaics. In order to evaluate the methods proposed in this thesis, two different pCLE databases were constructed, one on the colonic polyps and one on the Barrett's esophagus. Due to the initial lack of a ground truth for CBIR of pCLE, we first performed an indirect evaluation of the retrieval methods, using nearest-neighbor classification. Then, the generation of a sparse ground truth, containing the similarities perceived between videos by multiple experts in pCLE, allowed us to directly evaluate the retrieval methods, by measuring the correlation between the retrieval distance and the perceived similarity. Both indirect and direct retrieval evaluations demonstrate that, on the two pCLE databases, our retrieval method outperforms several state-of-the-art methods in CBIR. In terms of binary classification, our retrieval method is shown to be comparable to the offline diagnosis of human expert endoscopists on the *Colonic Polyps* database.

Because establishing a pCLE diagnosis is an everyday practice, our objective is not only to support one-shot diagnosis but also to accompany the endoscopists in their progress. Using retrieval results, we estimate the *difficulty* to interpret a pCLE video. We show that there is a correlation between the estimated difficulty and the diagnosis difficulty experienced by multiple endoscopists. The proposed difficulty estimator could thus be used in a self-training simulator, with difficulty level selection, which should help the endoscopists in shortening their learning curve.

The standard visual-word-based distance already provides adequate results for pCLE retrieval. Nevertheless, little clinical knowledge is embedded in this distance. By incorporating prior information about the similarity perceived by pCLE experts, we are able to learn an adjusted visual similarity distance which we prove to be better than the standard distance. In order to learn pCLE semantics, we then leverage multiple semantic concepts used by the endoscopists to describe pCLE videos. As a result, visual-word-based *semantic* signatures are built which extract, from low-level visual features, a higher-level clinical knowledge that is expressed in the endoscopist own language.

Keywords: probe-based Confocal Laser Endomicroscopy (pCLE); Gastrointestinal cancers; Content-Based Image Retrieval (CBIR); Bag-of-Visual-Words (BoW) method; Difficulty of pCLE video interpretation; Similarity distance learning; Semantic gap.



MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

Atlas Intelligent pour Guider le Diagnostic en Endomicroscopie : Une Application Clinique de la Reconnaissance d'Images par le Contenu

Résumé : L'Endomicroscopie Confocale par Minisondes (ECM) permet l'observation dynamique des tissus au niveau cellulaire, *in vivo et in situ*, pendant une endoscopie. Grâce à ce nouveau système d'imagerie, les médecins endoscopistes ont la possibilité de réaliser des "biopsies optiques" non invasives. Les biopsies traditionnelles impliquent le diagnostic *ex vivo* d'images histologiques par des médecins pathologistes. Le diagnostic *in vivo* d'images ECM est donc un véritable challenge pour les endoscopistes, qui ont en général seulement un peu d'expertise en anatomopathologie. Les images ECM sont néammoins de nouvelles images, qui ressemblent visuellement aux images histologiques. Cette thèse a pour but principal d'assister les endoscopistes dans l'interprétation *in vivo* des séquences d'images ECM.

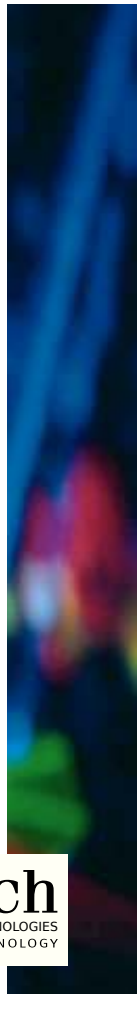
Lors de l'établissement d'un diagnostic, les médecins s'appuient sur un raisonnement par cas. Afin de mimer ce processus, nous explorons les méthodes de Reconnaissance d'Images par le Contenu (CBIR) pour l'aide au diagnostique. Notre premier objectif est le développement d'un système capable d'extraire de manière automatique un certain nombre de vidéos ECM qui sont visuellement *similaires* à la vidéo requête, mais qui ont en plus été annotées avec des métadonnées comme par exemple un diagnostic textuel. Un tel système de reconnaissance devrait aider les endoscopistes à prendre une décision éclairée, et par là-même, à établir un diagnostic ECM plus précis.

Pour atteindre notre but, nous étudions la méthode des Sacs de Mots Visuels, utilisée en vision par ordinateur. L'analyse des propriétés des données ECM nous conduit à ajuster la méthode standard. Nous mettons en œuvre la reconnaissance de vidéos ECM complètes, et pas seulement d'images ECM isolées, en représentant les vidéos par des ensembles de mosaïques. Afin d'évaluer les méthodes proposées dans cette thèse, deux bases de données ECM ont été construites, l'une sur les polypes du côlon, et l'autre sur l'œsophage de Barrett. En raison de l'absence initiale d'une vérité terrain sur le CBIR appliquée à l'ECM, nous avons d'abord réalisé des évaluations indirectes des méthodes de reconnaissance, au moyen d'une classification par plus proches voisins. La génération d'une vérité terrain éparse, contenant les similarités perçues entre des vidéos par des experts en ECM, nous a ensuite permis d'évaluer directement les méthodes de reconnaissance, en mesurant la corrélation entre la distance induite par la reconnaissance et la similarité perçue. Les deux évaluations, indirecte et directe, démontrent que, sur les deux bases de données ECM, notre méthode de reconnaissance surpasse plusieurs méthodes de l'état de l'art en CBIR. En termes de classification binaire, notre méthode de reconnaissance est comparable au diagnostic établi offline par des endoscopistes experts sur la base des *Polypes du Côlon*.

Parce que diagnostiquer des données ECM est une pratique de tous les jours, notre objectif n'est pas seulement d'apporter un support pour un diagnostique ponctuel, mais aussi d'accompagner les endoscopistes dans leurs progrès. A partir des résultats de la reconnaissance, nous estimons la *difficulté* d'interprétation des vidéos ECM. Nous montrons l'existence d'une corrélation entre la difficulté estimée et la difficulté de diagnostic éprouvée par plusieurs endoscopistes. Cet estimateur pourrait ainsi être utilisé dans un simulateur d'entraînement, avec différents niveaux de difficulté, qui devrait aider les endoscopistes à réduire leur courbe d'apprentissage.

La distance standard fondée sur les mots visuels donne des résultats adéquats pour la reconnaissance de données ECM. Cependant, peu de connaissance clinique est intégrée dans cette distance. En incorporant l'information *a priori* sur les similarités perçues par les experts en ECM, nous pouvons apprendre une distance de similarité qui s'avère être plus fidèle que la distance standard à la similarité perçue. Dans le but d'apprendre la sémantique des données ECM, nous tirons également profit de plusieurs concepts sémantiques utilisés par les endoscopistes pour décrire les vidéos ECM. Des signatures *sémantiques* fondées sur mots visuels sont alors construites, capables d'extraire, à partir de caractéristiques visuelles de bas niveau, des connaissances cliniques de haut niveau qui sont exprimées dans le propre language de l'endoscopiste.

Mots clés : Endomicroscopie Confocale par Minisondes (ECM) ; Cancers gastrointestinaux ; Reconnaissance d'images par le contenu ; Méthode des sacs de mots visuels ; Difficulté d'interprétation des vidéos ECM ; Apprentissage de la distance de similarité ; Fossé sémantique.



MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY