A Smart Atlas for Endomicroscopy using Automated Video Retrieval

Barbara André, Tom Vercauteren, Anna M. Buchner, Michael B. Wallace, Nicholas Ayache

A Smart Atlas for Endomicroscopy using Automated Video Retrieval

Barbara André [a,b,*], Tom Vercauteren [b], Anna M Buchner [c], Michael B. Wallace [d], Nicholas Ayache [a]

[a] *INRIA Sophia Antipolis, Asclepios Research Project, 2004 route des Lucioles - BP 93, 06902 Sophia Antipolis Cedex, France*

[b] *Mauna Kea Technologies,9 rue d'Enghien, 75010 Paris, France*

[c] *Hospital of the University of Pennsylvania, 3400 Spruce Street, PA, USA*

[d] *Mayo Clinic, 4500 San Pablo Road, Jacksonville, FL, USA*

Abstract

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. To remove 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 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 neighbors 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.

Key words: Content-Based Video Retrieval (CBVR), Endomicroscopy, Bag-of-Visual-Words (BoW), Video-Mosaicing

* Corresponding author.
Email addresses: barbara.andre@sophia.inria.fr (Barbara André),

1 Introduction

With the recent technology of probe-based confocal laser endomicroscpy (pCLE), physicians are able to image the epithelium at microscopic level with a miniprobe and in real time during an ongoing endoscopy procedure. As mentioned by Wallace and Fockens (Wallace and Fockens, 2009), the main task for the endoscopists is to establish a diagnosis from the acquired pCLE videos, by relating a given appearance of the epithelium to a specific pathology. They detect tissue areas that are suspicious for disease and either perform confirmatory biopsy, or if high certainty exists, perform immediate therapy such as resection or ablation of diseased tissue. Because standard endoscopic imaging can only diagnose disease states with moderate levels of certainty, biopsy is frequently performed, some of which are ultimately found to be normal tissue. Furthermore, the need for confirmatory biopsy delays a diagnosis and often requires a separate endoscopic procedure to be performed for treatment.

Currently, pCLE is relatively new to many physicians, who are still in the

tom.vercauteren@maunakeatech.com (Tom Vercauteren), annabuchner@hotmail.com (Anna M Buchner), Wallace.Michael@mayo.edu (Michael B. Wallace), nicholas.ayache@sophia.inria.fr (Nicholas Ayache).

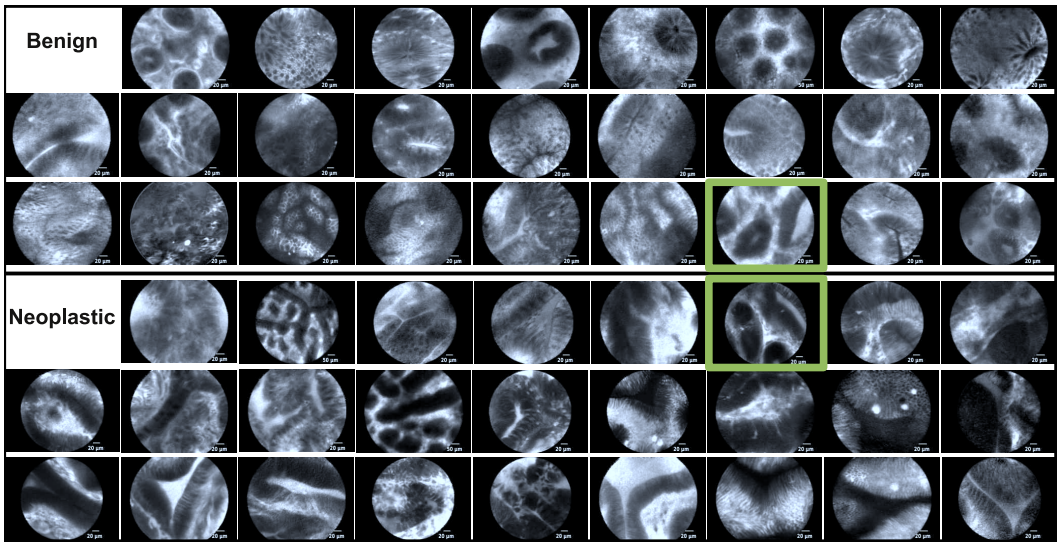

Fig. 1. **pCLE image samples from our database of colonic polyps.** The images have a diameter of approximately 500 pixels that corresponds to a FoV of 240 μm. Images of the polyps diagnosed as benign are on the top, whereas those diagnosed as neoplastic are on the bottom. The closer to the boundary 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.

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 method to evaluate retrieval performance is classification. In our approach, we make a clear distinction between retrieval, which is the target in this study, and classification, which is the indirect means that we choose to evaluate it. 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 by Müller et al. (Müller et al., 2004) in 2004. Particularly, the medical image retrieval task of ImageCLEF, presented in (Müller et al., 2008), 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, for example by Kong et al. (Kong et al., 2009), or by Doyle et al. (Doyle et al., 2006). Nevertheless, many standard computer-aided diagnosis criteria that are commonly employed in histology 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 3 the Bag-of-Visual-Words (BoW) method, proposed by Sivic and Zisserman (Sivic and Zisserman, 2006), which has been successfully used in many applications of computer vision. To apprehend the large intra-class variability of our pCLE database, we refer the reader to Fig. 1, 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 2.1, as well as the type of invariance necessary for their retrieval. By adjusting the image description to these invariants in Section 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, as proposed by Leung and Malik (Leung and Malik, 2001) 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 in Section 4 larger mosaic images that are built from the image sequences using the video-mosaicing technique of Vercauteren et al. (Vercauteren et al., 2006). The high degree of variability in appearance also holds for the resulting mosaic images, as shown in Fig. 12. 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. To reach this objective, in Section 5, we took advantage of the coarse registration results of real-time mosaicing to 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 accuracies 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 6, with encouraging results.

2 Context of the Study

2.1 Probe-based Confocal Laser Endomicroscopy

The principle of pCLE consists of inserting, through the standard endoscope, a miniprobe made of tens of thousands of optical fibers. As illustrated in Fig. 2, a laser scanning uses two mirrors to emit, along each fiber, an excita-

tion 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, endomicroscopic images are acquired at a rate of 12 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 et al., 2004) produces single images of diameter 500 pixels, which corresponds to a FoV of 240 μm, as illustrated in Fig. 5. All the pCLE video sequences that are used for this study have been acquired by the Cellvizio system of Mauna Kea Technologies. In stable video sequences the probe is in constant contact with the tissue, so the distance of the probe's optical center to the tissue is fixed.

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

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

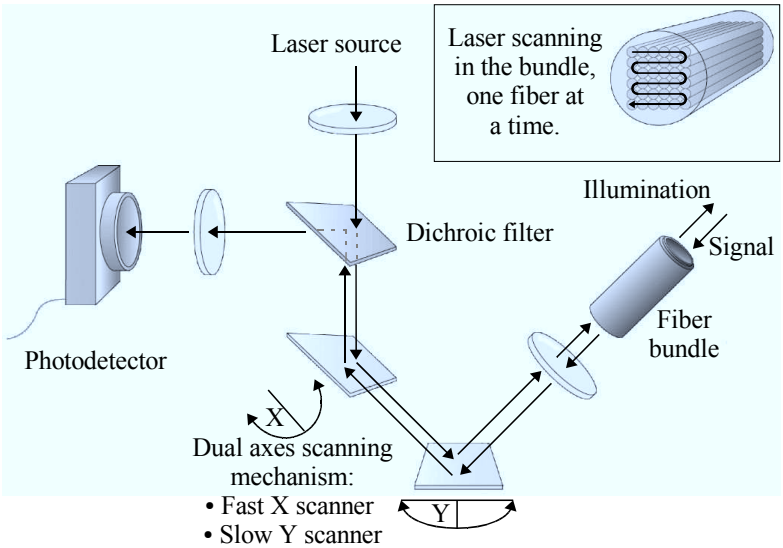

Fig. 2. **Principle of pCLE imaging.**

neoplastic tissue which contains some irregularities in the cellular and vascular architectures. The resulting 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 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 et al., 2008), (Buchner et al., 2009b), which included a video database of colonic polyps comparable to ours, and demonstrated the effectiveness of pCLE classification of polyps by experts endoscopists.

2.3 Framework for Retrieval Evaluation

Assessing the quality of content-based data retrieval is a difficult problem. In this paper, we focus on a simple 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 classes, benign and neoplastic, then we propose a multi-class evaluation of the retrieval in Section 6. As an objective indicator of the retrieval relevance, we take the classification accuracy (number of correctly classified samples / total number of samples).

It is worth mentioning that, in the framework of medical information retrieval, some scenarios require predefined sensitivity or specificity goals, depending on the application. 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 propose a Bayesian cost model for nearest-neighbors classification by introducing a weighting parameter θ to trade-off the cost of false positives and false negatives. This allows us to generate ROC curves as follows: when considering k nearest neighbors for a query, we compute the value of the weighted sum of their votes (-1 for benign class, $+1$ for neoplastic class) according to their similarity distance to the query, and we compare this value with the absolute threshold θ to classify the query as benign or neoplastic. The closer θ is 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. 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 parameter θ may also be useful to reduce the bias implied by our unbalanced dataset, which contains more benign images than pathological ones.

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 et al., 2004): 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 sub-sequences were done by one single expert. For our study we can consider this bias as negligible.

2.4 State-of-the-Art Methods in CBIR

In the field of computer vision, Smeulders et al. (Smeulders et al., 2000) presented a large review of the state-of-the-art in CBIR. In a closely related study, using an image database of colonic polyps but from a macroscopic point of view, Häfner et al. (Häfner et al., 2009) worked on endoscopic images and obtained rather good classification results by considering 6 pathological classes. However, their goal is classification for computer-aided diagnosis, whereas our main objective is retrieval. Petrou et al. (Petrou et al., 2006) proposed a solution for the description of irregularly-sampled images, which could be defined by the optical fiber positions in our case. Nevertheless, we will not work on irregularly-sampled images, but rather on the interpolated images. 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 et al., 2007) which is referred to as the HH-SIFT method combining sparse feature extraction with the BoW model, we will take as references the following methods for CBIR method comparison: the standard approach of Haralick features (Haralick, 1979) based on global statistical features and experimented by Srivastava et al. (Srivastava et al., 2008) in a closely related setup, the texture retrieval Textons method of Leung and Malik (Leung and Malik, 2001) based

on dense local features, but also an interesting image classification method presented by Boiman et al. (Boiman et al., 2008), the Naive-Bayes Nearest-Neighbor (NBNN) method. A brief description of these methods is provided in the supplemental material. 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 we decided to compare it with the NBNN method, that uses no clustering and that was proven to outperform BoW-based classifiers in (Boiman et al., 2008)

In order to determine if the improvement from one method to another is statistically significant, we will perform the McNemar's test (Sheskin, 2004) based on the classification results obtained by the two methods at a fixed number of nearest neighbors. The principle of the McNemar's test is explained in the supplemental material.

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

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 labels called "visual words", whose frequencies constitute the image signature. As illustrated in Fig. 3, 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 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 distance is that it is only based on the comparison between the values within the same histogram bin: if $H_1 = (v_1, ..., v_K)$ and $H_2 = (w_1, ..., w_K)$ are the histograms of the two images, then $\chi^2(H_1, H_2) = \frac{1}{2}\sum_{i=1}^{K}(v_i - w_i)^2/(v_i + w_i)$. In these conditions, as explained by Sivic and Zisserman (Sivic and Zisserman, 2006), 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 and Stewenius, 2006) showed that, combined with a hierarchical clustering, the inverted file indexing enables large-scale data retrieval. More sophisticated metrics, like the Earth Mover's Distance (EMD) proposed by Rubner et al. (Rubner et al., 2000), are less computationally efficient as they need to compute in the high-dimensional space the distances between the description vectors. Nevertheless, it would be interesting to test the fast implementation of EMD that has been recently presented by Pele and Werman (Pele and Werman, 2009). For the classification step that quantifies the similarity results, the votes of the k-nearest neighbors are weighted by the inverse of their χ^2 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, 2004). 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 et al., 2007) 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. 6: when considering only 2 classes, the accuracy is below 67%, which is not compatible with 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

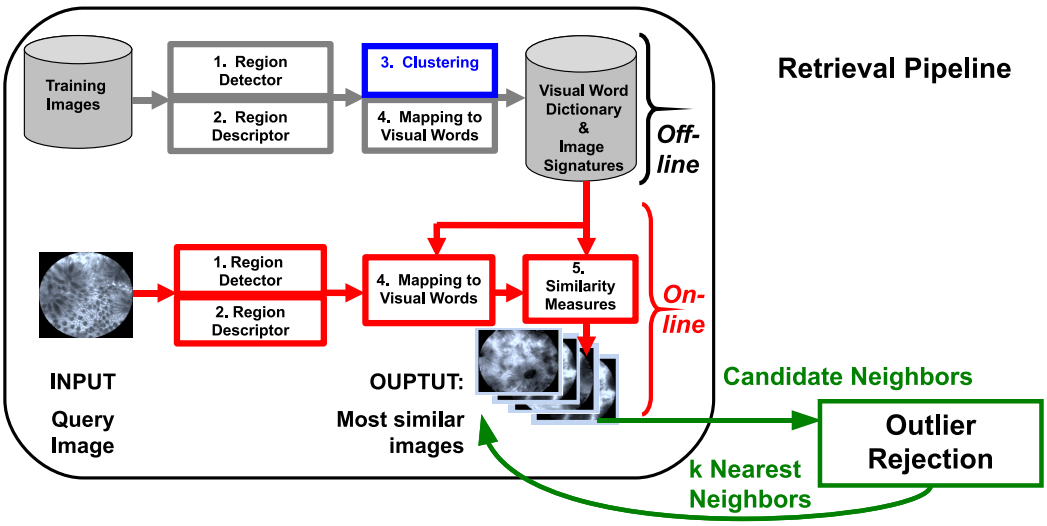

Fig. 3. **Overview of the retrieval pipeline, potentially followed by a the geometrical verification process.**

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 3 and 4. We explored them in a preliminary study (André et al., 2009a).

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. 5. 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 and Van Gool, 2000) and the Maximally Stable Extremal Regions (MSER) of Matas et al. (Matas et al., 2002) are also specialized for the extraction of blob features.

However, while testing on pCLE videos the numerous sparse detectors listed in (Mikolajczyk et al., 2005), 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. 4. 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. 5. 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 et al., 2006), who were mainly interested in object categorization and segmentation problems.

3.3 Multi-Scale Description of Local Regions

Let us now look at what kinds 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 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 et al., 2007) 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 et al., 2006) 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

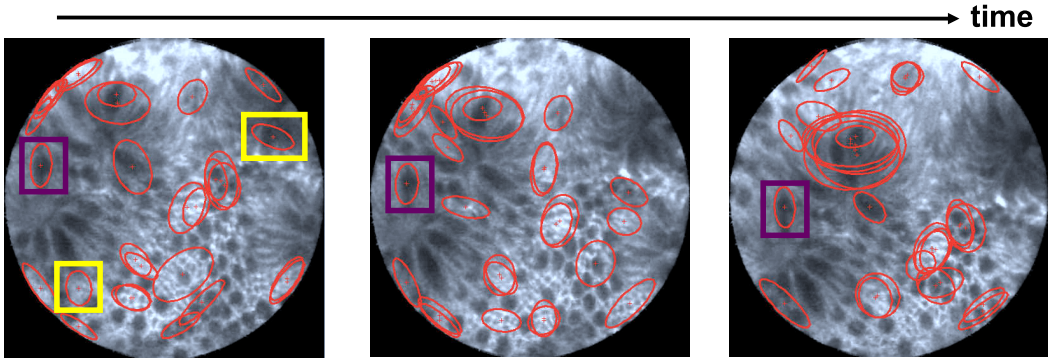

Fig. 4. **Salient regions (ellipses) extracted by the sparse MSER detector on three successive frames of a benign video sequences.** Some regions, like the one framed in dark, are correctly followed by the detector, but many others, like those framed in bright, are lost. This shows the inconsistency of the sparse detector for the description of pCLE images.

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. 6 where we observe that, compared to the standard HH-SIFT method, the dense detector brings a gain of accuracy of 17.1 percentage points (p.p.) 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 $k \in [1, 10]$).

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. 5. For the classification of single images, we observe in Fig. 6 that, when compared to the one-scale description of the Dense-Scale-60 (D-S-60) method, the bi-

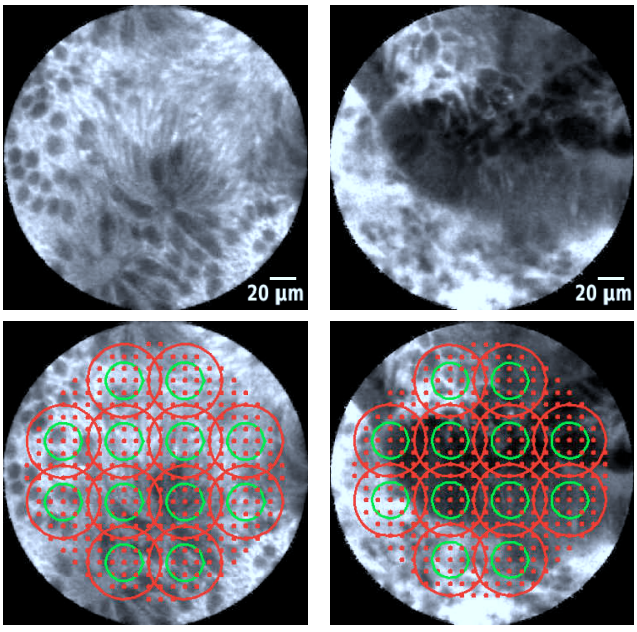

Fig. 5. **Small and large disk regions on a dense regular grid, applied on a benign image (left), and on a neoplastic image (right).** Small disks of radius 30 pixels cover microscopic information like individual cells, whereas large disks of radius 60 pixels cover mesoscopic information like groups of cells. The images have a diameter of approximately 500 pixels that corresponds to a FoV of 240 μm.

scale description of the Dense-Bi-Scale-30-60 (D-BS-30-60) method brings an additional gain of accuracy of 2.5 p.p. 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.

4 Contributions to the State-of-the-Art

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. 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 et al., 2006), to project the temporal dimension of a video sequence onto one mosaic image with a larger FoV

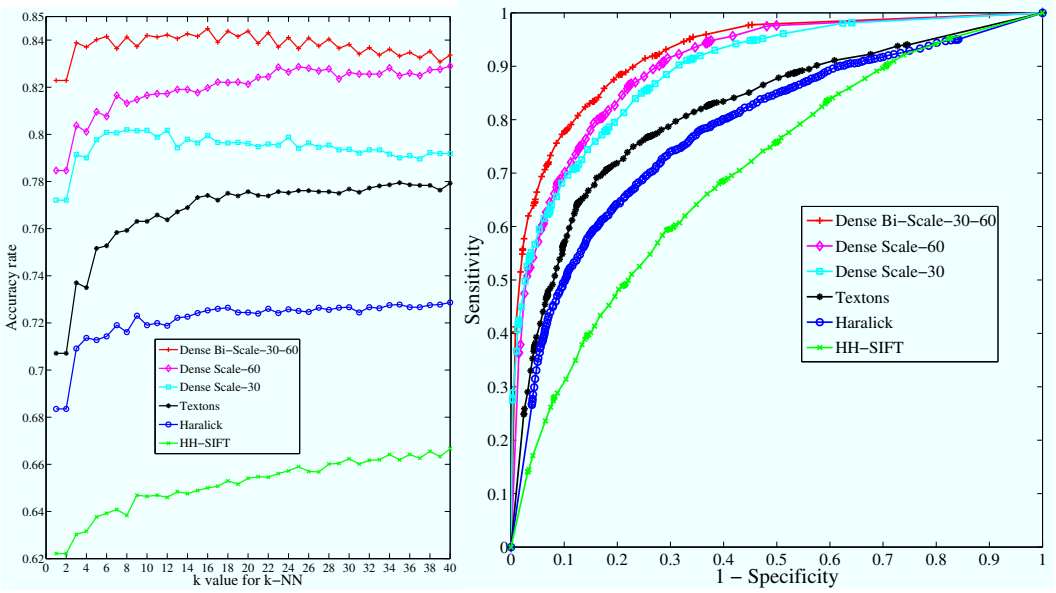

Fig. 6. **Left: LOPO classification of single pCLE images by the methods, with $\theta = 0$. Right: Corresponding ROC curves at $k = 10$ neighbors with $\theta \in [-1, 1]$.** θ trades off the cost of false positives and false negatives.

and of higher resolution. Even if time information is lost after the mosaicing, Becker et al. (Becker et al., 2007) showed that the mosaic image produced by this video-mosaicing technique has a clinical interest in endomicroscopy.

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. 7. 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. More investigation is needed to understand the causes of this observation. 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.

4.2 Similarity Metric based on Visual Words

The similarity metric defined by the χ^2 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 of images, we considered the distribution $p(w|C)$ of the number of occurrences of a visual word w in the images belonging to the class C. The discriminative power $f(w)$ of the visual word w is chosen by using the Fisher criterion which can be expressed as the Mahalanobis distance between the two distributions $p(w|C_1)$ and $p(w|C_2)$: $f(w) = (\mu_1 - \mu_2)^2/(0.5\ (\sigma_1^2 + \sigma_2^2))$, where μ_i and σ_i^2 are respectively the mean and the variance of the distribution of w in the images belonging to class i. Our approach, that combines $L1$-normalization applied to the visual word histograms, and Fisher weighting applied to the visual words, could be composed with other similarity metrics

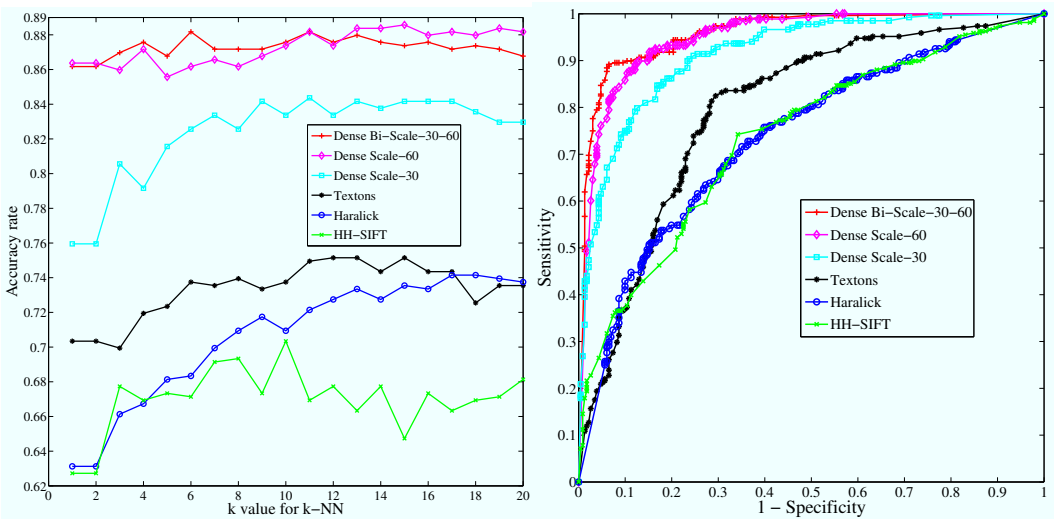

Fig. 7. **Left: LOPO classification of pCLE mosaic images by the methods, with $\theta = 0$. Right: Corresponding ROC curves at $k = 5$ neighbors with $\theta \in [-1, 1]$.** θ trades off the cost of false positives and false negatives. The mosaic images have been built with non-rigid registration.

than χ^2, some of which are presented in (Sivic and Zisserman, 2009). 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 and Dance, 2007) as an extension of the BoW method for image categorization. In particular, a binary weighting leads to the selection of the most discriminative visual words, i.e. those minimizing the intra-class distances while maximizing the inter-class distances. Furthermore, by reducing the number of visual words, the size of image signatures is decreased, 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' for all cross-validation sub-sets is 23.2.

The classification of mosaic images presented in Fig. 8 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 p.p. 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 $k = 3$): it brings an additional gain of accuracy of 3.4 p.p. at $k = 5$ neighbors, with a resulting accuracy of 90.2% (93.7% sensitivity, 86.2% specificity).

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. Jegou et al. (Jegou et al., 2008) previously proposed to add a geometrical verification that takes spatial information into account. However their method is 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, we do not have any quantitative measure of perceived similarity. 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 took advantage of the dense distribution of visual words to define their adjacency using the 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. 9. We investigated this idea in a prior study (André et al., 2009b). 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

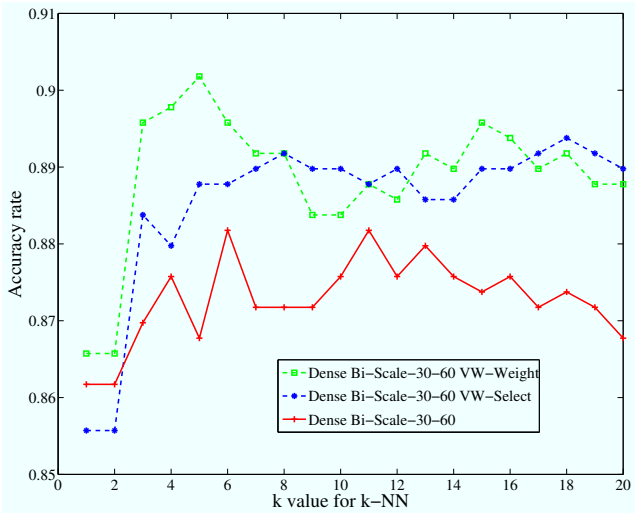

Fig. 8. **LOPO classification of pCLE mosaic images (with $\theta = 0$) using the discriminative power of the visual words.** The mosaic images have been built with non-rigid registration.

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 are stored in a l x 1 dimensional vector denoted by $\mathbf{m}$, where l is equal to the number of the lower triangular elements. The LDA weights, represented as a l x 1 dimensional vector denoted by $\mathbf{L}$, satisfy: $\mathbf{L} = \mathbf{\Sigma^{-1}} (\mu_1 - \mu_2)$, 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 μ_i is the mean of the vector $\mathbf{m}$ associated with all the training images belonging to the class i. 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. 3.

In practice, to prevent from over-fitting on our database, the number of LDA

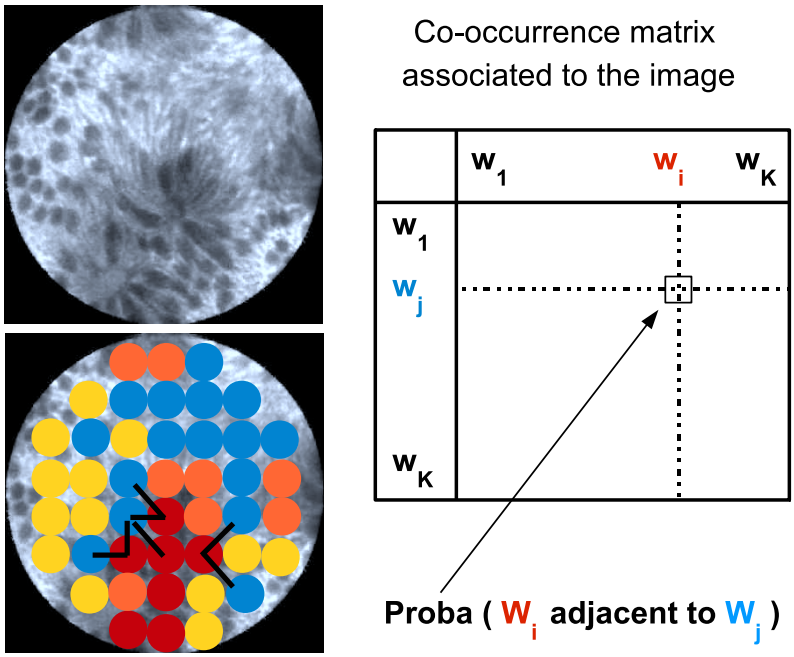

Fig. 9. **Example of a co-occurrence matrix M associated to a benign image.** M is a symmetric matrix of size $K \times K$ where K is the number of visual words. Considering 2 visual words, respectively associated to the colors blue and red, black edges link the blue-labeled regions and the red-labeled 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.

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. 10 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 p.p., with a resulting accuracy of 83.2% (82.8% sensitivity, 83.6% specificity). Besides, when added to the one-scale description with disks of radius 60 pixels, the outlier removal brings an additional gain of accuracy, even though 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.

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. Potential ways of doing so include their incorporation into the description as proposed by Zhang et al. (Zhang et al., 2009), or their extraction at hierarchical scales in the image as described in the Hyperfeatures of Agarwal and Triggs (Agarwal and Triggs, 2008).

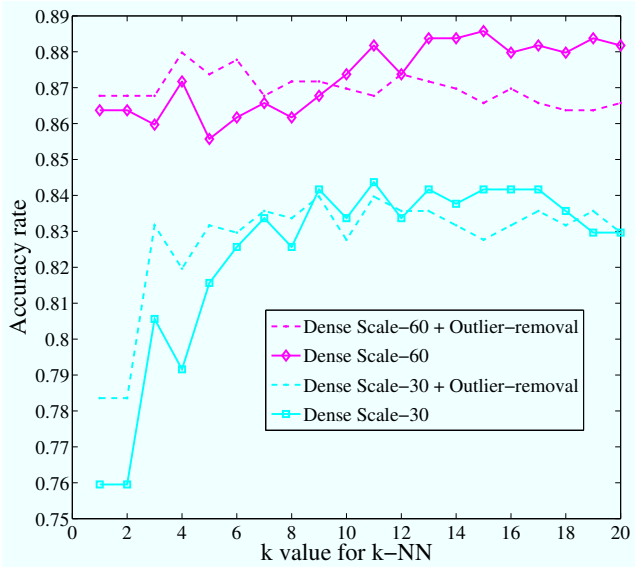

Fig. 10. **LOPO classification of pCLE mosaic images (with $\theta = 0$) using outlier removal.** The mosaic images have been built with non-rigid registration.

5 Endomicroscopic Videos Retrieval using Implicit Mosaics

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 et al., 2010), 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 during the image acquisition time.

For this purpose, we first compute independently the signatures of 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. 11 on the right. To define the signature H 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. $w(.)$ 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: $H_S(i) = \frac{1}{Z} \sum_{I \in S} \sum_{r \in I} \delta(w(r), i)/\Gamma(r)$. 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: $Z = \sum_{i \in [1,K]} \sum_{I \in S} \sum_{r \in I} \delta(w(r), i)/\Gamma(r)$.

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. 13, 14, 15, 16, 17 and with more examples in the supplemental material.

5.2 Method Comparison for Video Retrieval

Our methodological improvements, from image retrieval to video retrieval, depend on several conditions: the used techniques, i.e. overlap weighting and histogram summation, but also the objects of interest for the retrieval, i.e. single images, fused mosaic images, video sub-sequences or full videos. In order to evaluate these improvements, 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. This is always the case excepted for the 121 full videos for which statistical significance cannot be tested (Sheskin, 2004). 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

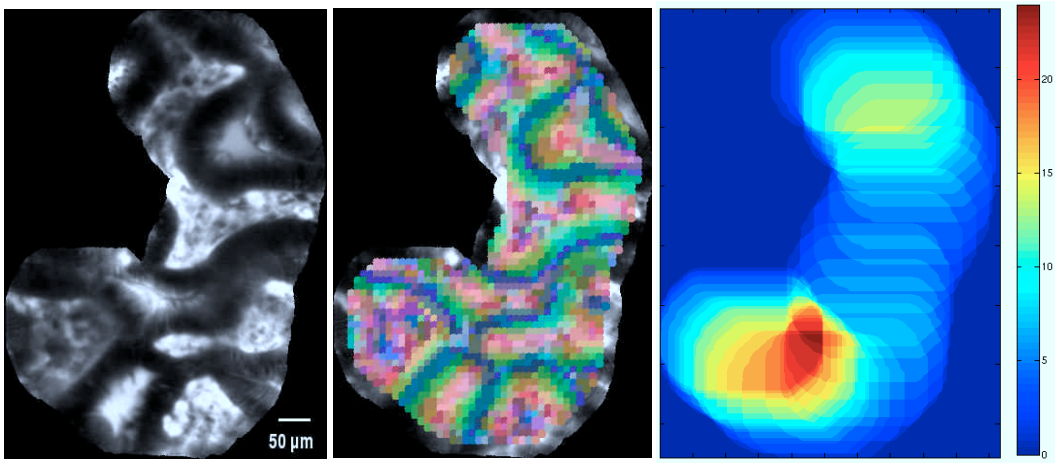

Fig. 11. **From left to right: Neoplastic pCLE mosaic obtained with non–rigid registration; Colored visual words mapped to the disk regions of radius** 60 **pixels in the mosaic image** (see the supplemental material for details on the coloring scheme) **; Overlap scores of the local regions in the mosaic space, according the translation results of mosaicing.**

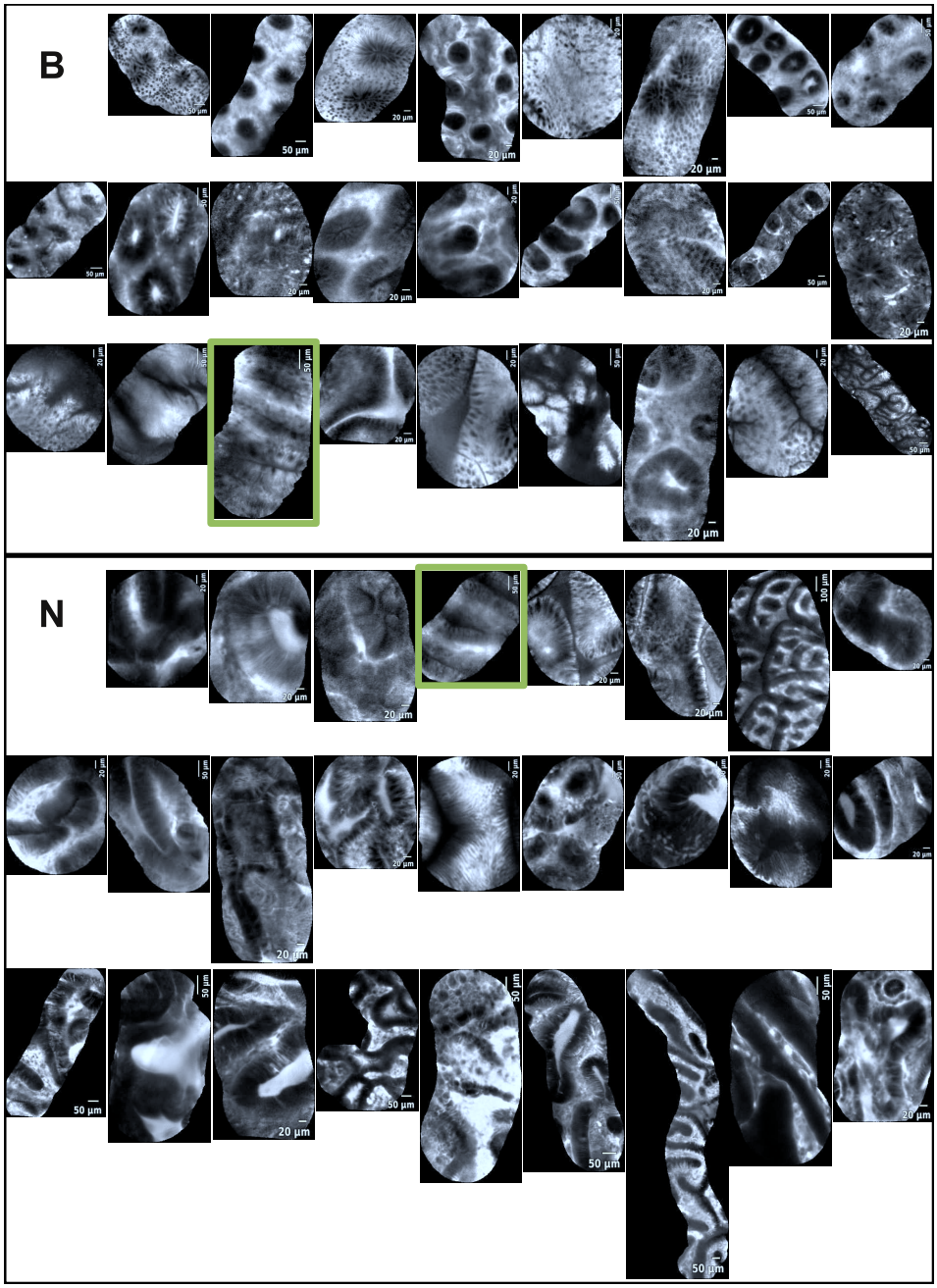

Fig. 12. **pCLE mosaic samples from our database of colonic polyps.** These mosaics have been built from image sequences using a video-mosaicing technique with non-rigid registration (Vercauteren et al., 2006). For visualization purposes, the size of the mosaics are not normalized. Mosaics of the polyps diagnosed as benign are on the top, indicated by **B**, whereas mosaics of the polyps diagnosed as neoplastic are on the bottom, indicated by **N**. The closer to the boundary the mosaics are, the less obvious is their diagnosis according to their visual appearance. In particular, the two framed mosaics 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.

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 4.1 describing the single fused mosaic image obtained with non-rigid registration; and "AverageVote-Im" the method describing all the images independently and

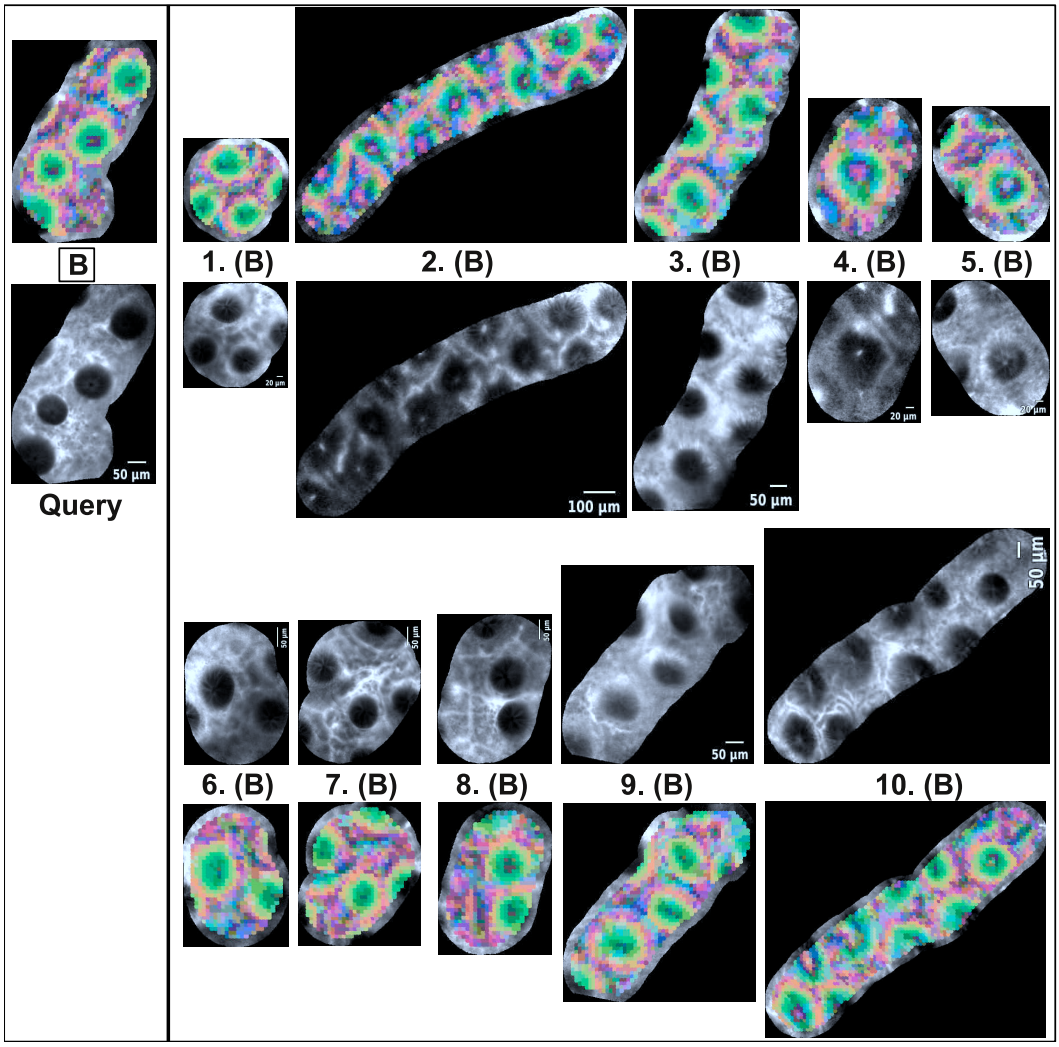

Fig. 13. **The** 10 **most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method**. The pCLE video sub-sequences are represented by their corresponding fused mosaic image built with non-rigid registration, that are shown together with their 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. The details on the coloring scheme for the visual words are explained in supplemental material. As a result, these colors are highlighting the geometrical structures in the mosaic images.

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.

When comparing the methods for the classification of video sub-sequences, Fig. 18 shows that the accuracy of "Weighted-ImOfMos" is better than the one of "AverageVote-Im", with statistical significance (p-value < 0.021 for $k \in [3, 10]$). For the classification of full videos, Fig. 19 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-

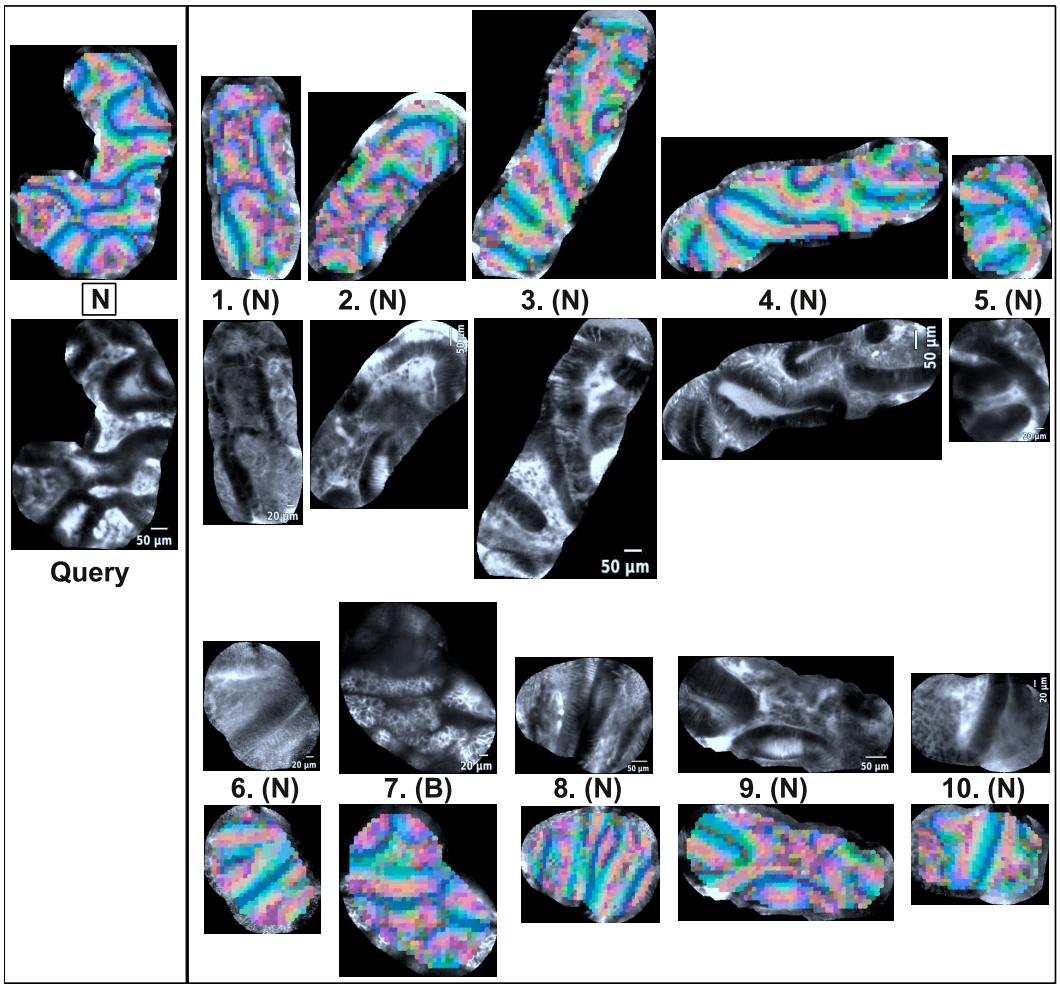

Fig. 14. **The 10 most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign (not present here) and **N** Neoplastic

.

Weighted-ImOfMos" already achieves a quite satisfying accuracy, e.g. 93.4% for 3 neighbors. Besides, for each method and for a fixed number of neighbors, a peak of classification accuracy is reached at a θ value which is more likely negative, as illustrated in the supplemental material for the "Weighted-ImOfMos" method with a slight accuracy peak at $\theta = -0.17$. This reflects the fact that neoplastic features are more discriminative than the benign ones.

For method comparison, we also tested an efficient classification method, the NBNN classifier of Boiman et al. (Boiman et al., 2008), which was mentioned in the Introduction. Although NBNN classifies images, we can easily extend it to a "Weighted-NBNN" method for the classification of video sub-sequences or full videos, by weighting the closest distance computed for each region by the inverse of its overlap score. Besides, a ROC curve for the "Weighted-NBNN" method can be obtained by introducing a multiplicative threshold θ_{Boi}, and by

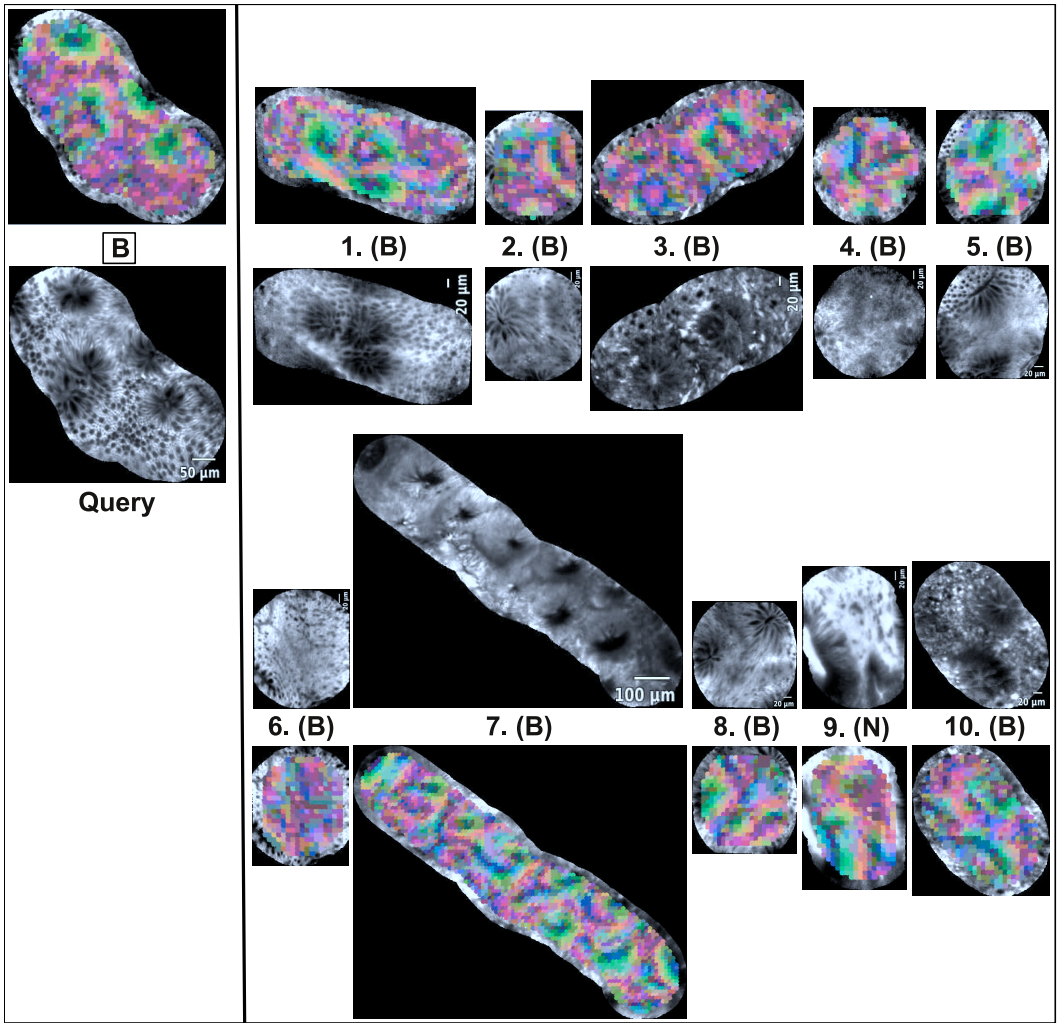

Fig. 15. **The 10 most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method. B** indicates Benign and **N** Neoplastic.

classifying the query as neoplastic if and only if $D_B < \theta_{Boi} D_N$, where D_B (resp. D_N) is the sum of the benign (resp. neoplastic distances) in the NBNN classifier. In comparison to the other methods, these ROC curves show worse results in Figs. 18 and 19, with statistical significance for the classification of video sub-sequences (p-values ≤ 0.05). Besides, the best classification accuracies of video sub-sequences by "Weighted-NBNN" are reached for $\theta_{Boi} = 0.98 < 1$, which is also confirming that local neoplastic features are more discriminative than the benign ones. 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.

6 Finer Evaluation of the Retrieval

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

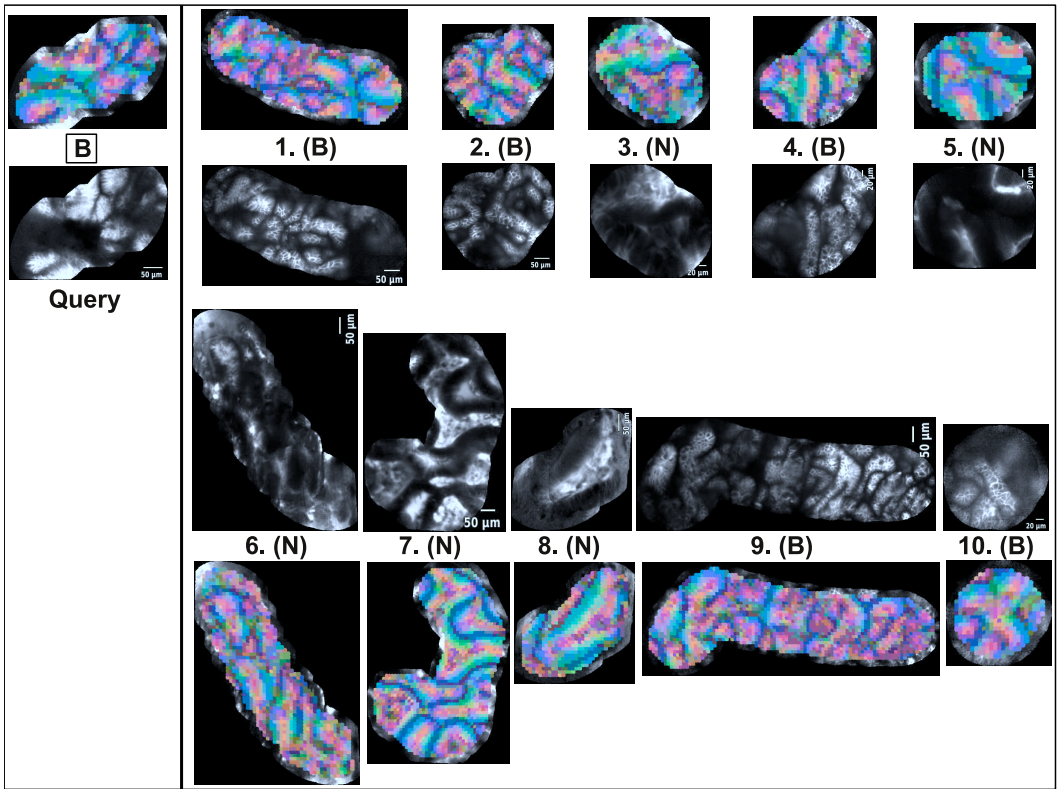

Fig. 16. **The 10 most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method. B** indicates Benign and **N** Neoplastic.

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

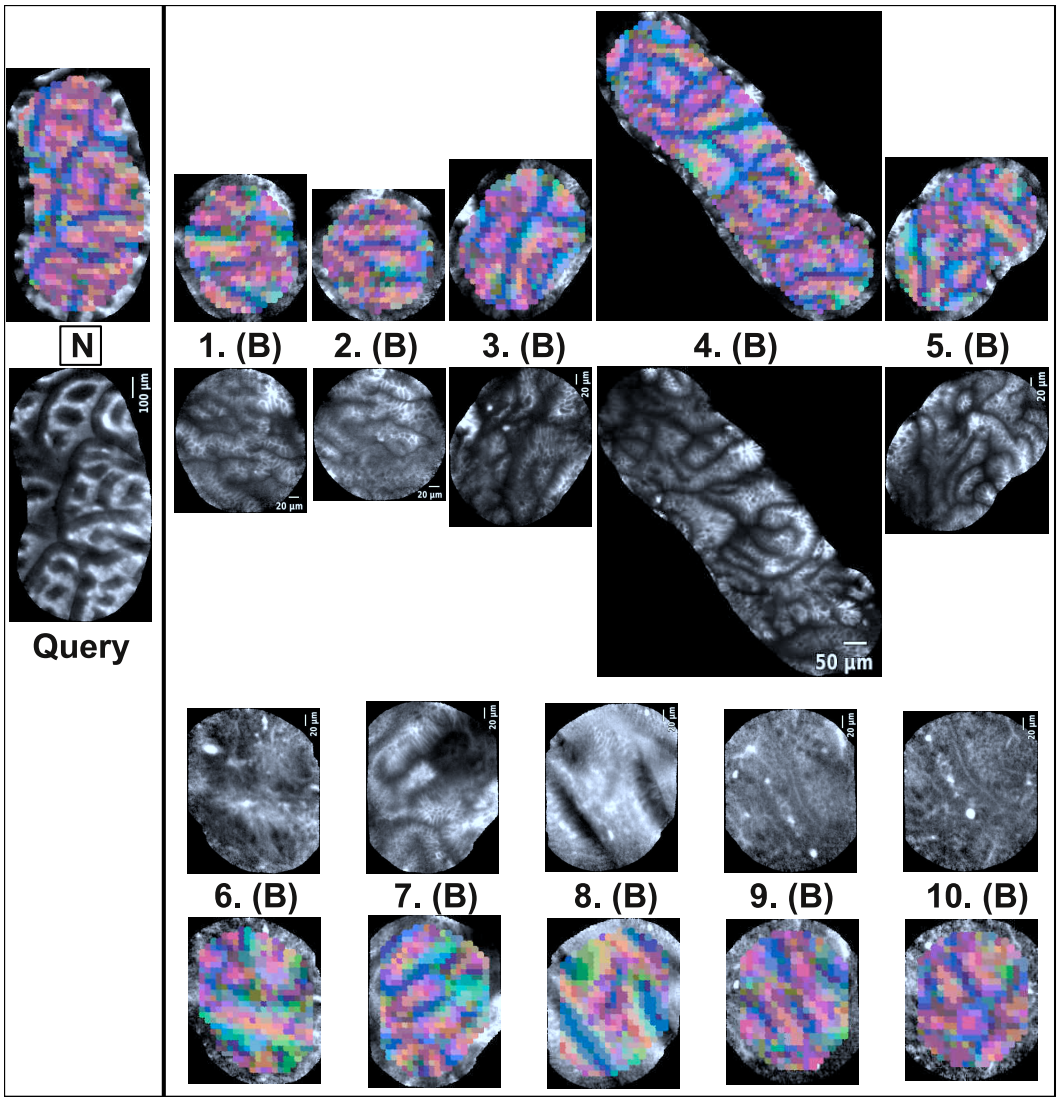

Fig. 17. **The 10 most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign and **N** Neoplastic. This query is a rare variety of the neoplastic class. This is one of the worst retrieval results, that are due to the relatively small size and weak representativity of the training database.

adenoma" (15 videos) and "adenocarcinoma" (4 videos).

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

Based on the finer diagnosis ground-truth, we perform a 5-class k-NN 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. To be able to measure a statistical significance, we take as objects of interest mosaic images instead of videos,

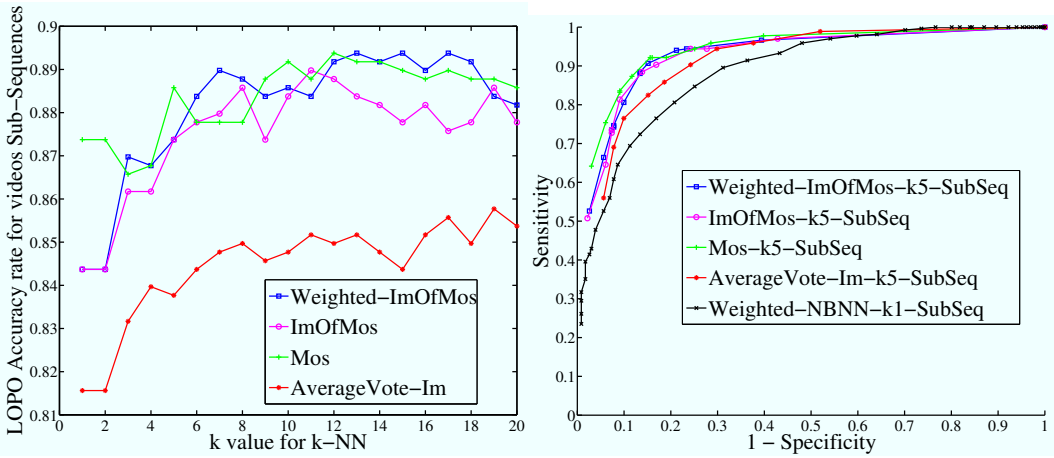

Fig. 18. **Left: LOPO classification of pCLE video sub-sequences, with $\theta = 0$. Right: Corresponding ROC curves at $k = 5$ neighbors with $\theta \in [-1, 1]$ and $\theta_{Boi} \in [1.0, 1.1]$.** θ trades off the cost of false positives and false negatives.

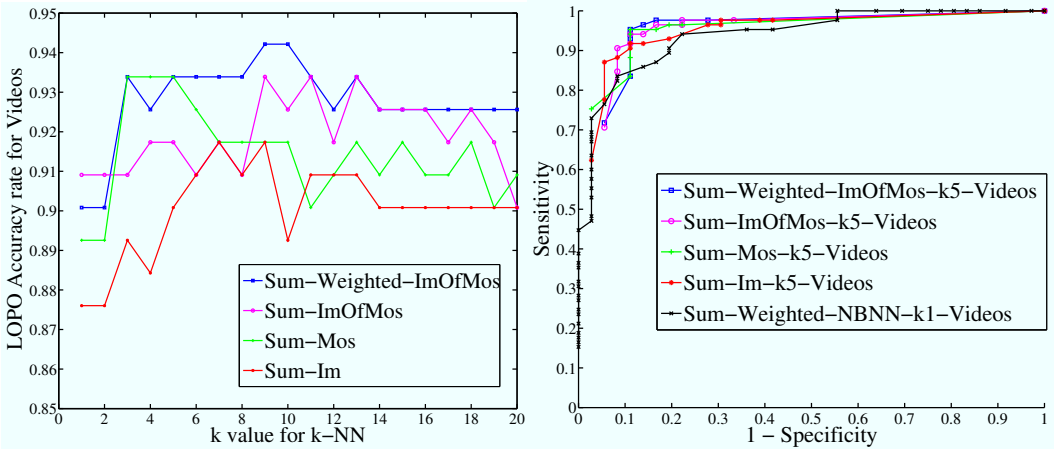

Fig. 19. **Left: LOPO classification of full pCLE videos, with $\theta = 0$. Right: Corresponding ROC curves at $k = 5$ neighbors with $\theta \in [-1, 1]$ and $\theta_{Boi} \in [1.0, 1.1]$.** θ trades off the cost of false positives and false negatives.

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. 20 and 21. 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 class 3 reaches an overall accuracy of 41.3 %, but 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 methods, with statistical significance from 3 nearest neighbors.

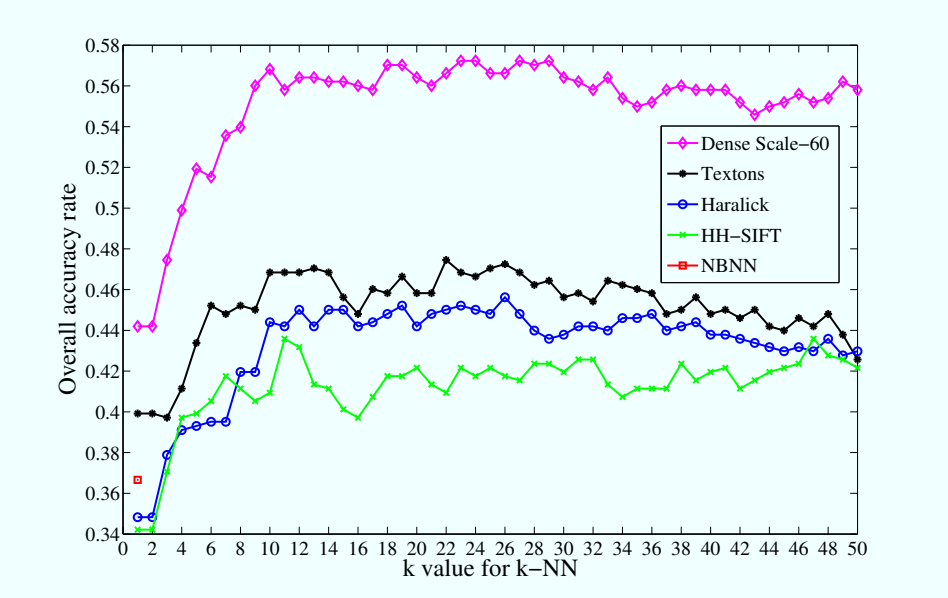

Fig. 20. **5-class LOPO classification of pCLE mosaic images by the methods.** The NBNN classification accuracy is represented at k=1, as it does not depend on k. The mosaic images have been built with non-rigid registration.

Method	Dense-Scale-60 $k = 10$	Textons $k = 10$	Haralick $k = 10$	HH-SIFT $k = 10$	NBNN $k = 1$
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 $k \geq 1$

Fig. 21. **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.

7 Conclusion

To 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, which potentially supports diagnostic decision and avoids unnecessary polypectomies of non-neoplastic lesions.

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 long-term goal is to generate a perceptual similarity ground-truth and directly evaluate the retrieval.

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 et al., 2008) already demonstrated the interest of endomicroscopy.

For future work, a larger training database would not only improve the classification results if all the characteristics of the image classes are better represented, but also enable the exploitation of the whole co-occurrence matrix of visual words at several scales. 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; it could incorporate as well the spatial information of multi-scale co-occurrence matrices into descriptors. 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 et al., 1999). 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. We could for example introduce spatio-temporal features, as it has been done by Wang et al. (Wang et al., 2009). This would allow us to work on more accurate visual words and better combine spatial and temporal information.

To conclude, the binary classification results that we obtained on our colonic polyp database compare favourably 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 et al., 2009a) showed an interobserver agreement with an average accuracy of 72% (sensitivity 82%, specificity 53%). Considering 3 expert endoscopists, Gomez et al. (Gomez et al., 2010) 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.

Acknowledgments

The authors would like to thank Dr. Muhammad Waseem Shahid who supported us in the construction of the endomicroscopic database of colonic polyps at the Mayo Clinic of Jacksonville, and Dr. Aymeric Perchant for his contribution to the progress of this study.

References

A. Agarwal and B. Triggs. Multilevel image coding with hyperfeatures. *Int. J. Comput. Vis.*, 78(1):15–27, 2008.

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 *Proc. ISBI'09*, pages 346–349, 2009a.

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)*, Sept. 2009b.

B. André, T. Vercauteren, A. Perchant, M. B. Wallace, A. M. Buchner, and N. Ayache. Endomicroscopic video retrieval using mosaicing and visual words. In *Proc. ISBI'10*, 2010.

H. Bay, T. Tuytelaars, and L. J. Van Gool. SURF: Speeded Up Robust Features. In *Proc. ECCV'06*, pages 404–417, 2006.

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. *Gastrointest. Endosc.*, 66(5):1001–1007, Nov. 2007.

O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Proc. CVPR'08*, pages 1–8, 2008.

A. M. Buchner, M. S. Ghabril, M. Krishna, H. C. Wolfsen, and M. B. Wallace. High-resolution confocal endomicroscopy probe system for in vivo diagnosis of colorectal neoplasia. *Gastroenterology*, 135(1):295, July 2008.

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*, 69(5):AB364–AB365, Apr. 2009a.

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. Wolfsen, and M. B. Wallace. Comparison of probe based confocal laser endomicroscopy with virtual chromoendoscopy for classification of colon polyps. *Gastroenterology*, in press, Nov. 2009b.

X. Descombes, R. Morris, J. Zerubia, and M. Berthod. Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood. *IEEE Trans. Image Process.*, 8(7):954–963, July 1999.

S. Doyle, A. Madabhushi, M. D. Feldman, and J. E. Tomaszeweski. A boosting cascade for automated detection of prostate cancer from digitized histology. In *Proc. MICCAI'06*, pages 504–511, 2006.

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 *Int. J. Comput. Assisted Radiol. Surg.*, pages 1010–1014, 2004.

V. Gomez, A. M. Buchner, E. Dekker, F. J. van den Broek, A. Meining, M. W. Shahid, M. Ghabril, P. Fockens, and M. B. Wallace. Interobserver agreement and accuracy among international experts of probe-based confocal laser microscopy (pCLE) in predicting colorectal neoplasia. *Endoscopy*, in press, 2010.

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 *Proc. MICCAI'09*, pages 247–254, 2009.

R. M. Haralick. Statistical and structural approaches to texture. In *Proc. IEEE*, volume 67, pages 786–804, 1979.

H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV'08*, volume I, pages 304–317, Oct. 2008.

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 Recognit.*, 42 (6):1080–1092, 2009.

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 *Proc. MICCAI'04*, pages 761–768, 2004.

T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vis.*, 43:29–44, June 2001.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60:91–110, Nov. 2004.

J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. British Mach. Vision Conf.*, 2002.

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. *Int. J. Comput. Vis.*, 65:43–72, Nov. 2005.

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. *I. J. Medical Informatics*, 73(1):1–23, 2004.

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.

D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR'06*, pages 2161–2168, 2006.

O. Pele and M. Werman. Fast and robust earth mover's distances. In *Proc. ICCV'09*, 2009.

F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. *Proc. CVPR'07*, pages 1–8, 2007.

M. Petrou, R. Piroddi, and A. Talepbour. Texture recognition from sparsely and irregularly sampled data. *Comput. Vis. Image Underst.*, 102(1):95–104, 2006.

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*, 57(12):1648–1653, 2008.

Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40(2):99–121, Nov. 2000.

D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, Boca Raton, Florida, USA, third edition, 2004.

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 *Proc. ECCV'06*, pages 1–15, 2006.

J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*, pages 127–144, 2006.

J. Sivic and A. Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):591–606, 2009.

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 Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

S. Srivastava, J. J. Rodriguez, A. R. Rouse, M. A. Brewer, and A. F. Gmitro. Computer-aided identification of ovarian cancer in confocal microendoscope images. *J. Biomed. Opt.*, 13(2):024021, March/April 2008.

T. Tuytelaars and L. J. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proc. British Mach. Vision Conf.*, 2000.

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. *Med. Image Anal.*, 10(5):673–692, Oct. 2006.

M. B. Wallace and P. Fockens. Probe-based confocal laser endomicroscopy. *Gastroenterology*, 136(5):1509–1513, May 2009.

H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. British Mach. Vision Conf.*, page 127, Sept. 2009.

J. Zhang, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *Int. J. Comput. Vis.*, 73:213–238, June 2007.

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.

8 Supplemental Material

8.1 Technical Details

8.1.1 Details on the State-of-the-Art Methods

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 3.

The Naive-Bayes Nearest-Neighbor (NBNN) classifier of (Boiman et al., 2008) is a simple yet efficient image classification method that uses no clustering. For each local region of the query it 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.

8.1.2 McNemar's Test as Statistical Relevance Measure

Let $n_{1,2}$ be the number of query objects that are correctly classified by the first method and misclassified by the second method. Let $n_{2,1}$ be the number of query objects that are correctly classified by the second method and misclassified by the first method. Then the McNemar's test statistic is given by: $\chi^2 = (|n_{1,2} - n_{2,1}| - 1)^2/(n_{1,2} + n_{2,1})$. Under the null hypothesis, if the observed differences are large enough ($n_{1,2} + n_{2,1} > 20$), then χ^2 has a chi-squared distribution with one degree of freedom and the associated p-value provides the statistical relevance.

8.1.3 Mapping Visual Words to Colors

To visualize the visual words in each described image, we decided 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 label. 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 $K = 100$ 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. 13. Besides, 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 to all the patients of the database.

8.2 Additional Figures

Here are the additional Figs. 22, 23, 24, 25, 26 and 27.

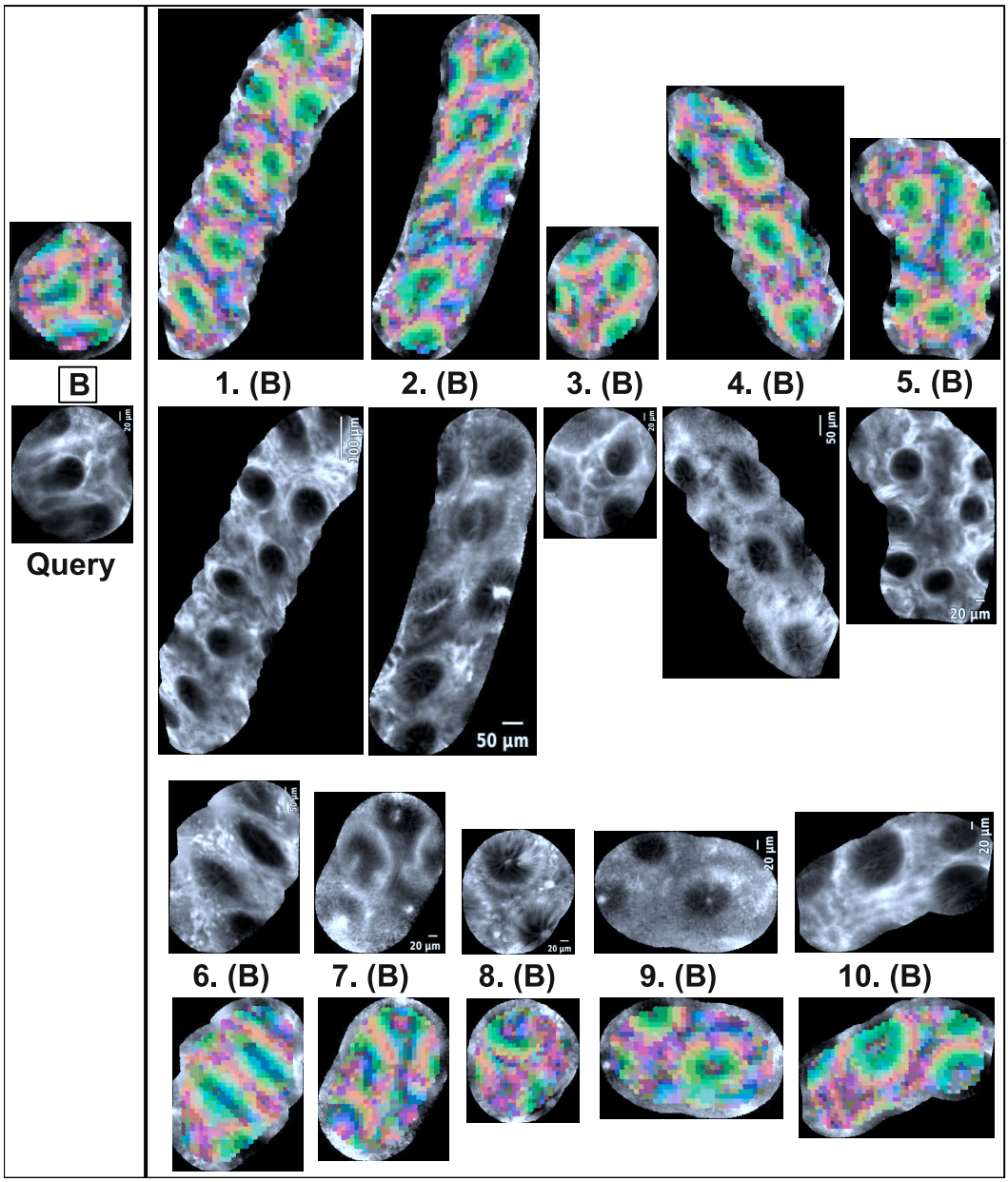

Fig. 22. **The** 10 **most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method**. The pCLE video sub-sequences are represented by their corresponding fused mosaic image built with non-rigid registration. **B** indicates Benign and **N** Neoplastic (not present here). For visualization purposes, the displayed visual words have been computed on the mosaic image on disks of radius 60 pixels. The details on the coloring scheme for the visual words are explained in supplemental material. As a result, these colors are highlighting the geometrical structures in the mosaic images.

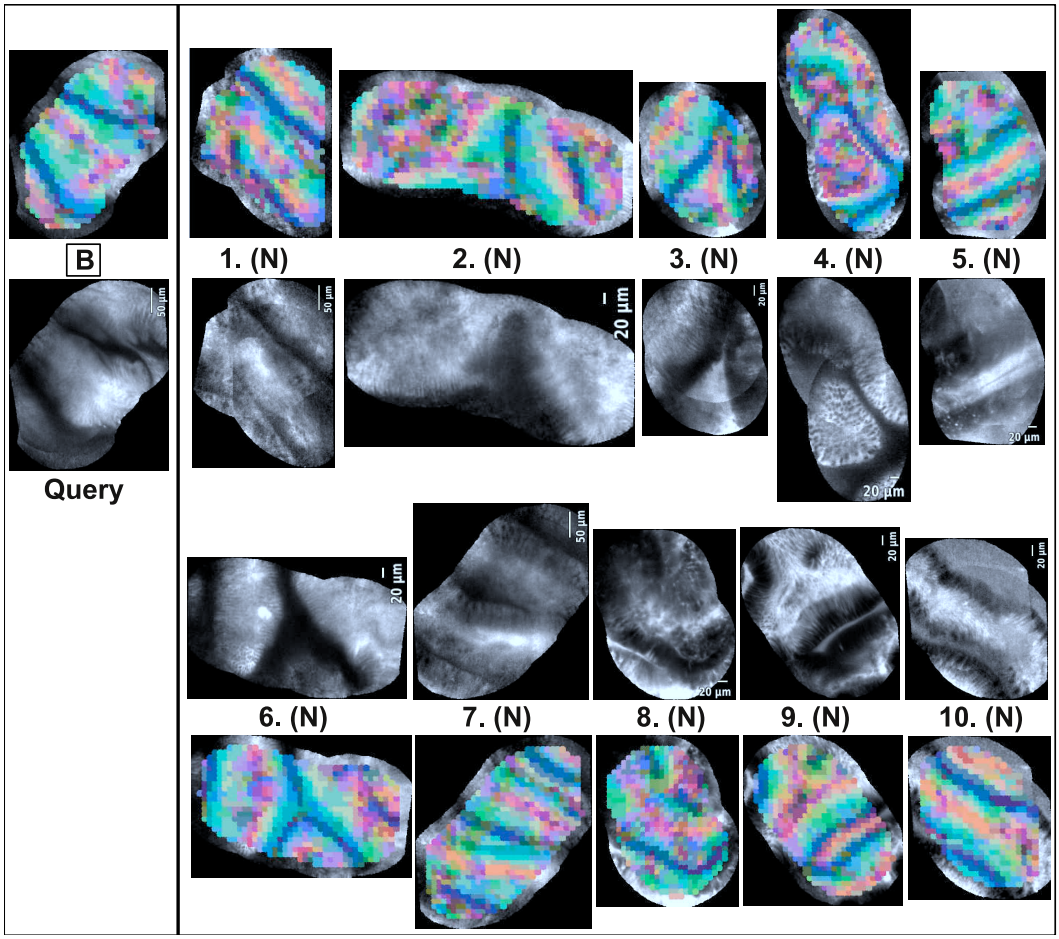

Fig. 23. **The** 10 **most similar pCLE video sub-sequences (right) for a benign query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign and **N** Neoplastic. Such bad retrieval result appears when the query is a rare variety of its pathological class, and is explained by the relatively small size and weak representativity of the training database.

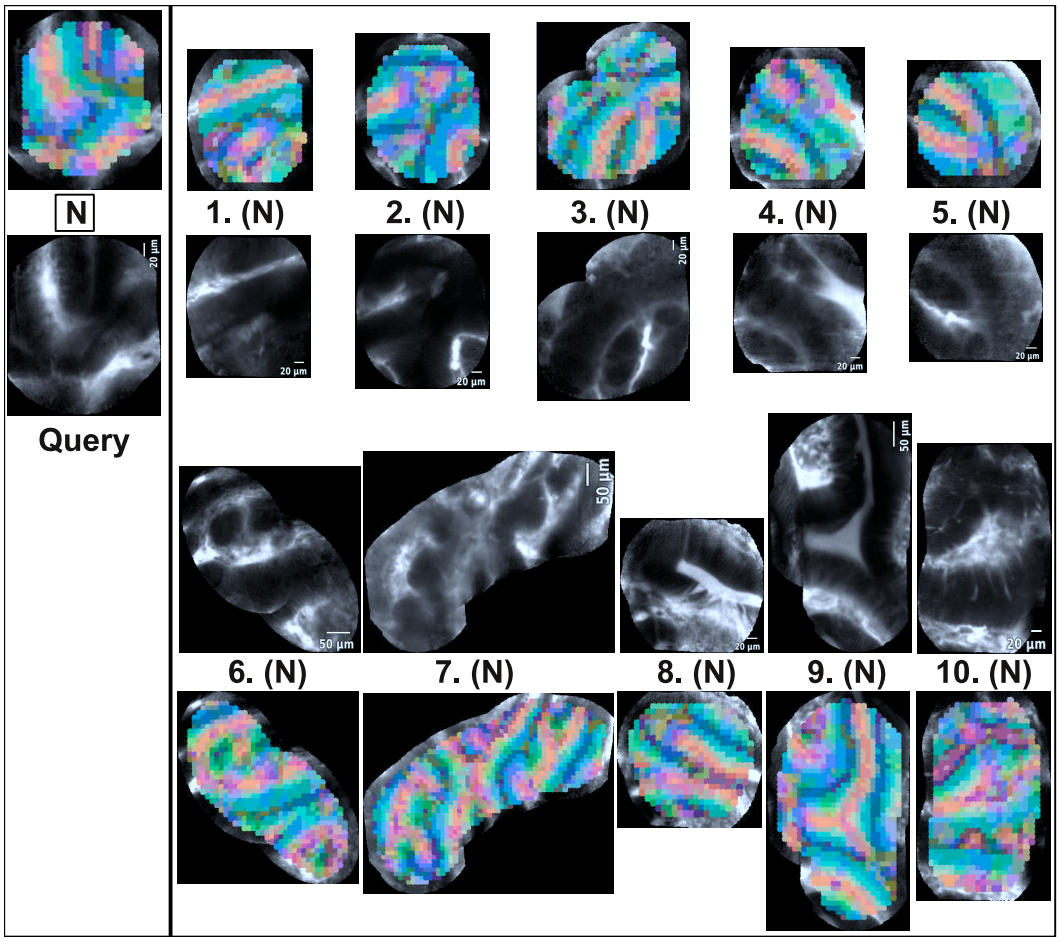

Fig. 24. **The** 10 **most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign (not present here) and **N** Neoplastic.

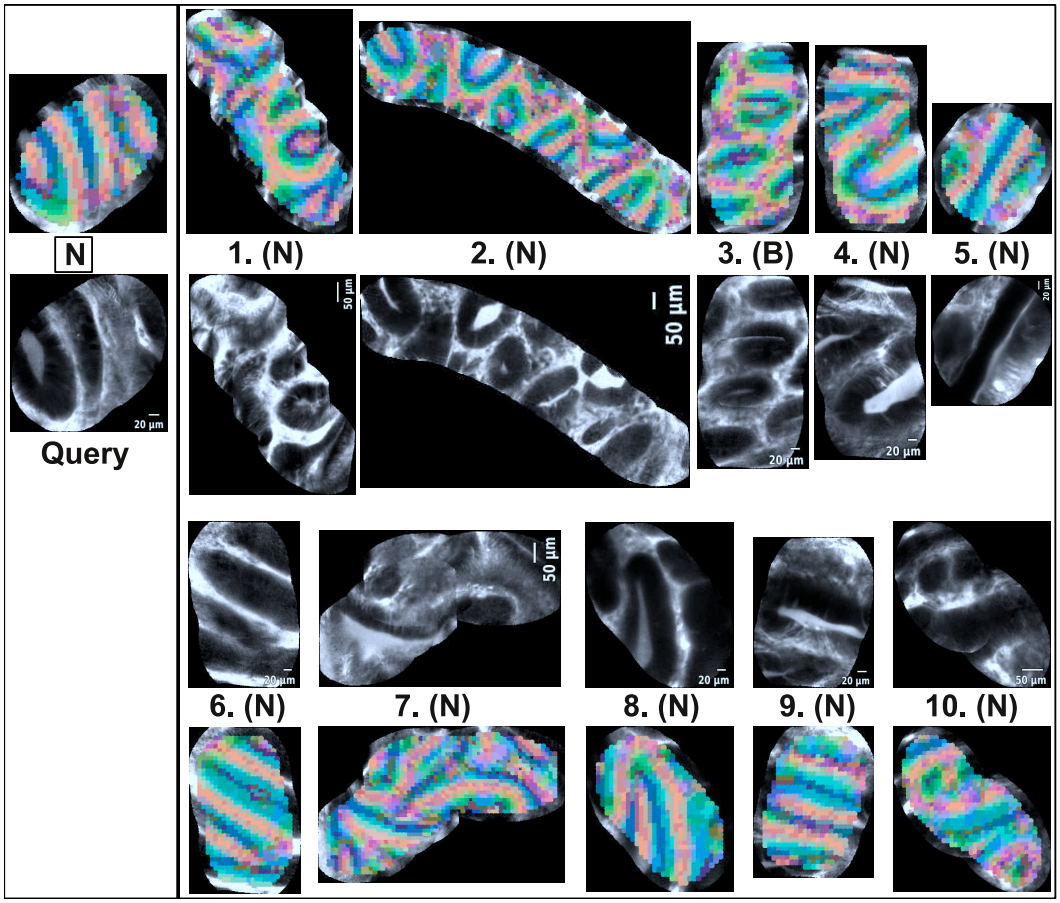

Fig. 25. **The** 10 **most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign and **N** Neoplastic.

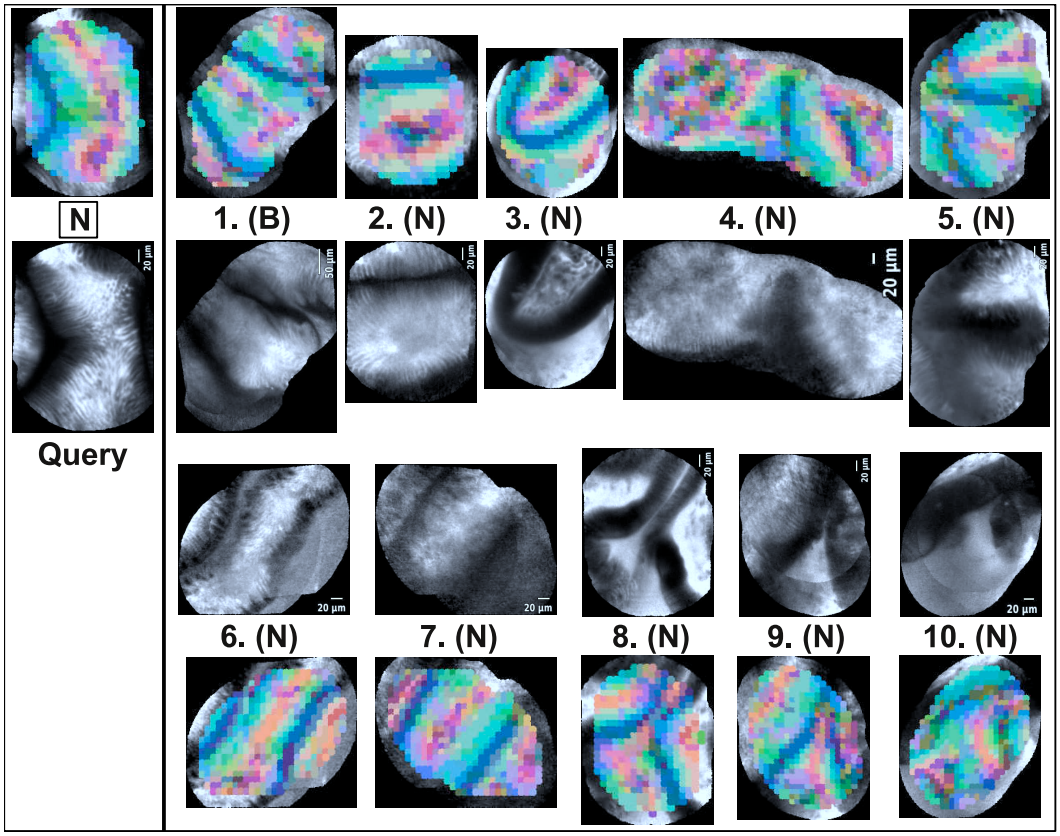

Fig. 26. **The 10 most similar pCLE video sub-sequences (right) for a neoplastic query (left), retrieved by the LOPO Weighted-ImOfMos method**. **B** indicates Benign and **N** Neoplastic.

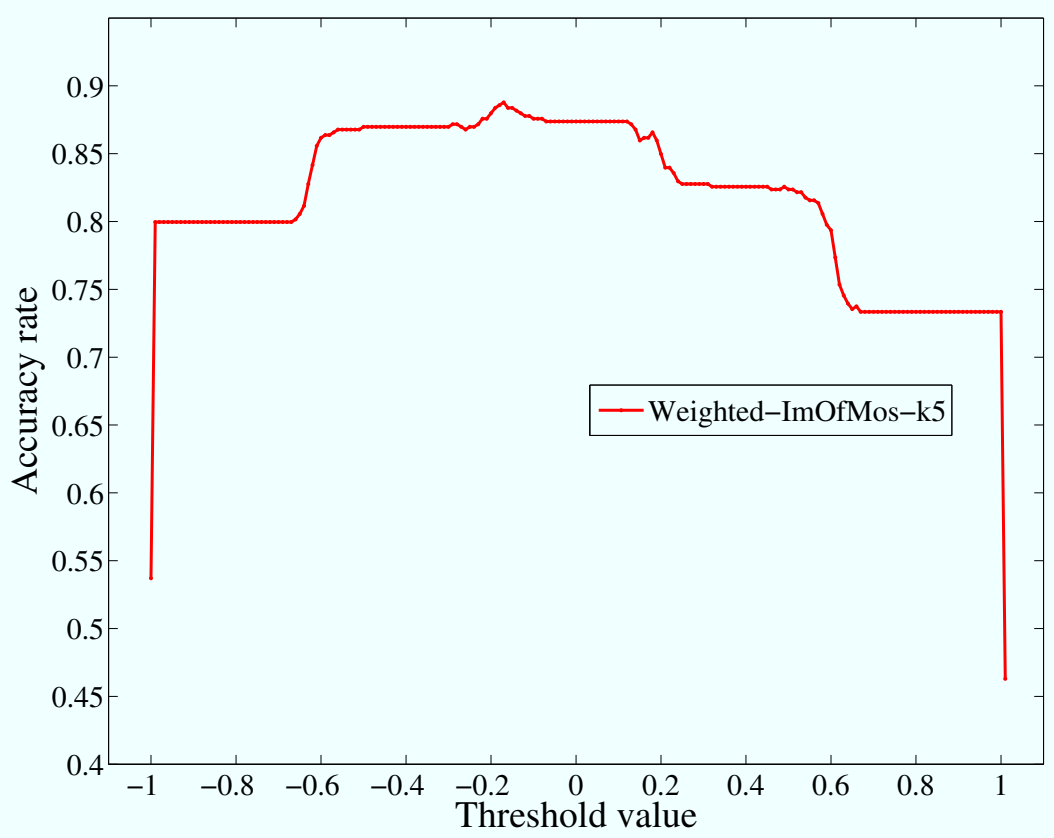

Fig. 27. **Accuracy rate for the classification of pCLE video sub-sequences by the LOPO Weighted-ImOfMos method, at $k = 5$ neighbors, depending on the value of the weighting threshold $\theta \in [-1, 1]$ that trades off the cost of false positives and false negatives.** The slight accuracy peak at the negative value $\theta = -0.17$ reflects the fact that neoplastic features are more discriminative than the benign ones.