Content-Based Retrieval in Endomicroscopy: Toward an Efficient Smart Atlas for Clinical Diagnosis

Barbara André[1,2] and Tom Vercauteren[1] and Nicholas Ayache[2]

[1]Mauna Kea Technologies (MKT), Paris, France
[2]INRIA - Asclepios Research Project, Sophia Antipolis

Abstract. In this paper we present the first Content-Based Image Retrieval (CBIR) framework in the field of *in vivo* endomicroscopy, with applications ranging from training support to diagnosis support. We propose to adjust the standard Bag-of-Visual-Words method for the retrieval of endomicroscopic videos. Retrieval performance is evaluated both indirectly from a classification point-of-view, and directly with respect to a perceived similarity ground truth. The proposed method significantly outperforms, on two different endomicroscopy databases, several state-of-the-art methods in CBIR. With the aim of building a self-training simulator, we use retrieval results to estimate the interpretation *difficulty* experienced by the endoscopists. Finally, by incorporating clinical knowledge about perceived similarity and endomicroscopy semantics, we are able: 1) to learn an adequate visual similarity distance and 2) to build visual-word-based *semantic* signatures that extract, from low-level visual features, a higher-level clinical knowledge expressed in the endoscopist own language.

1 Introduction

What is pCLE? Probe-based Confocal Laser Endomicroscopy (pCLE) allows the endoscopists to image the epithelium at a microscopic scale, *in vivo* and *in situ*, at real-time frame rate. Thanks to this novel imaging system illustrated in Fig. 1, the endoscopists have the opportunity to perform non-invasive *optical biopsies*. Traditional biopsies result in histological images that are classically diagnosed *ex vivo* by pathologists and not by endoscopists. The *in vivo* diagnosis of pCLE images is therefore a critical challenge for the endoscopists who typically have less expertise in histopathology. Fig. 2 illustrates this challenge by showing the high variability in appearance of pCLE mosaics of colonic polyps. Therefore, our main goal is to assist the endoscopists in the *in vivo* interpretation of pCLE image sequences.

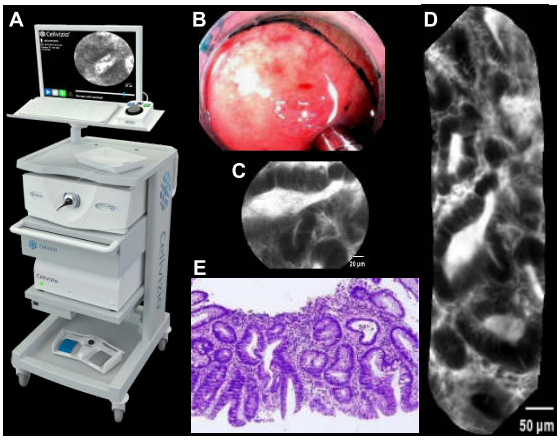

Fig. 1. (A) Setup of pCLE imaging system. (B) Endoscopic image of a colonic polyp diagnosed with tubular adenoma (macroscopic view), and the pCLE miniprobe. (C) Acquired pCLE image of the colonic polyp (microscopic "en-face" view, i.e. frontal view). (D) pCLE mosaic image associated to the acquired pCLE video. (E) Histopathology image of the colonic polyp (microscopic transverse view), obtained from a traditional biopsy corresponding to the "optical biopsy" site.

Why using CBIR to support pCLE diagnosis? 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 main objective is to develop a system which automatically extracts, from a training database, 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, acting like a *Smart Atlas* that opens a comprehensive book at the right pages, should help the endoscopist in making an informed decision and therefore a more accurate pCLE diagnosis. Another relevant application of our retrieval framework, which is presented in this paper, is to build a self-training simulator that features difficulty selection, in order to help the endoscopists in shortening their learning curve in pCLE diagnosis.

2 Adjusting Bag of Visual Words for pCLE Retrieval

Toward a Dense Bag-of-Visual-Words Method The Bag-of-Visual-Words method, proposed by Sivic and Zisserman [1], is a CBIR method which has been successfully applied in computer vision, in particular by Zhang et al. [2] for the classification of texture images. Another CBIR method based on multi-scale affine kernels is proposed by Syeda-Mahmood et al. in [3], with the purpose of object categorization. Noticing that pCLE images have a similar appearance to texture images and that their discriminative information is densely distributed, we propose in [4] a dense Bag-of-Visual-Words method for pCLE retrieval. Our

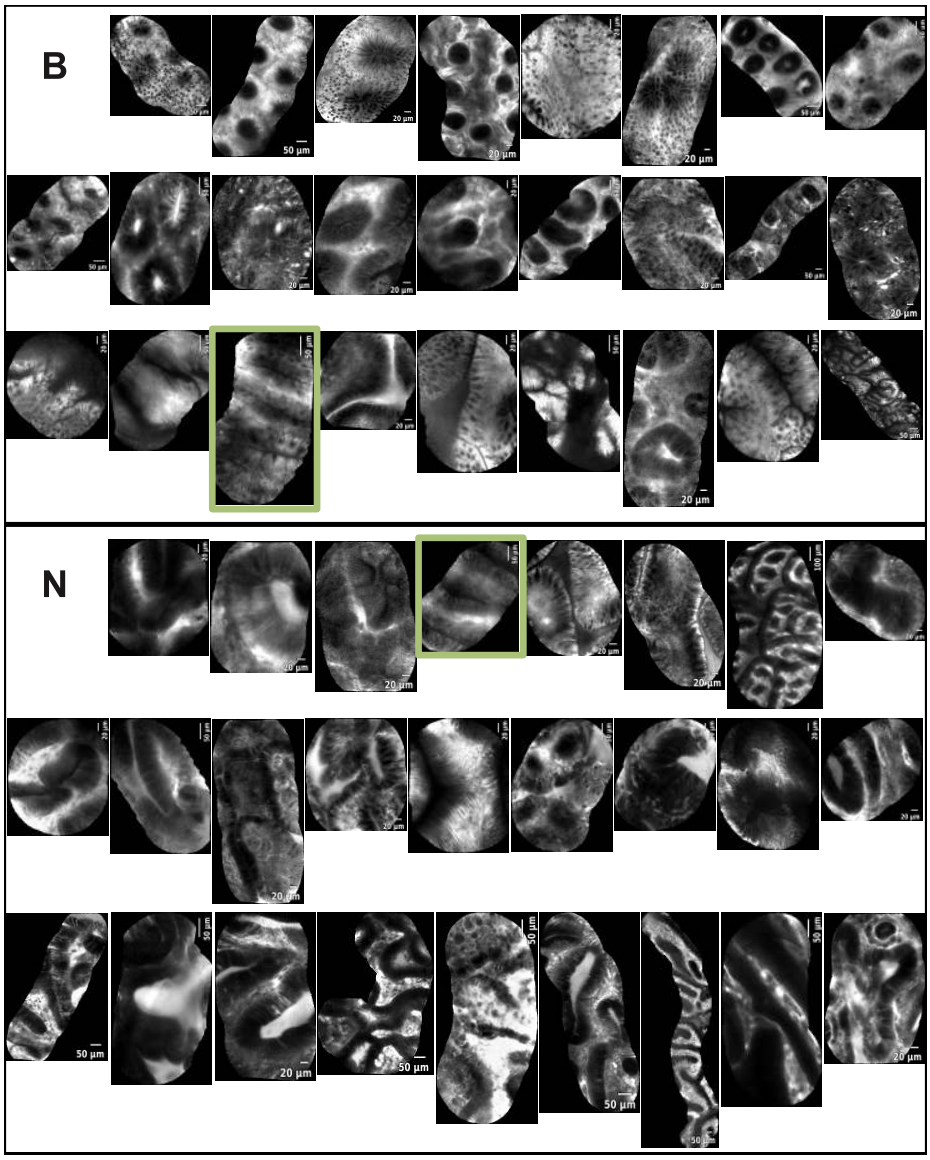

Fig. 2. Illustration of the *Semantic Gap* between low-level visual features and high-level clinical knowledge, on pCLE mosaics of the colonic polyps database. Scale bars provide a cue on the field of view size. On top (resp. bottom) are the mosaics of the polyps diagnosed as non-neoplastic (resp. neoplastic) 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.

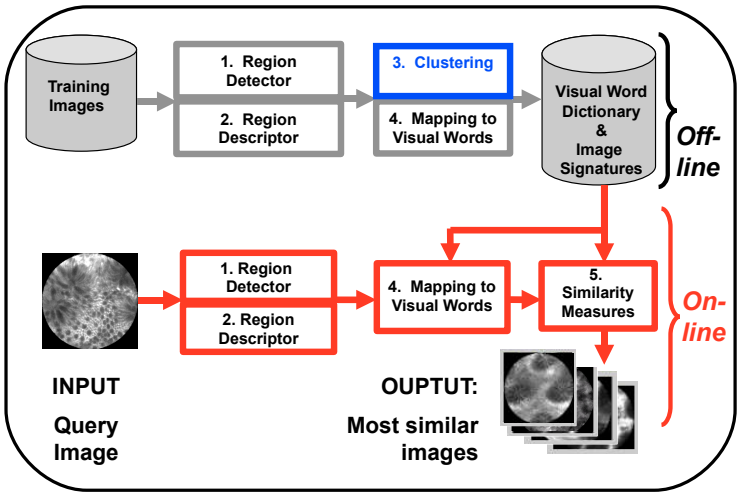

Fig. 3. Overview of the retrieval pipeline.

dense Bag-of-Visual-Words method consists of: 1) the detection of disk regions in the images at each point of a dense regular grid; 2) the description of these regions into description vectors using the Scale Invariant Feature Transform (SIFT); 3) the K-means clustering of all description vectors into K visual words; 4) the construction for each image of a visual word histogram, called *visual* signature; that is invariant with respect to viewpoint changes (translations and rotations) and illumination changes (affine transformations of the intensity). The whole retrieval pipeline is illustrated in Fig. 3.

Video Retrieval using Explicit or Implicit Mosaics As the pCLE miniprobe moves in constant contact with the tissue, the successive images from the acquired pCLE video are mostly related by viewpoint changes. We thus use the video mosaicing technique of Vercauteren et al. [5] that employs non-rigid registration to project the temporal dimension of a video sequence onto one mosaic image with a larger field of view and of higher resolution. Some examples of the resulting mosaic images are presented in Fig. 2. In [4], we propose two different representations of pCLE videos depending on the time constraints. In the *explicit mosaic* representation, mosaic images are built with the time-consuming non-rigid registration, then we compute the *visual* signature for each mosaic image, finally the video signature is obtained using a histogram summation technique. In the *implicit mosaic* representation, we first compute the *visual* signature for each single image, then we leverage coarse registration results between time-related images, provided by the real-time version of video mosaicing, in order to apply overlap weighting to the visual words before performing the histogram summation step. We define the similarity distance between two pCLE videos as the χ^2 pseudo-distance between their signatures. The resulting pCLE video

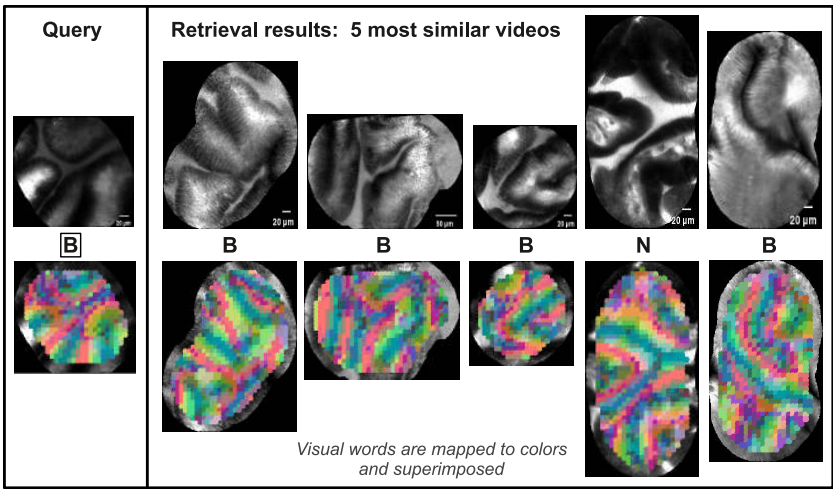

Fig. 4. Typical retrieval results on the Barrett's esophagus. The video query on the left is followed by its $k = 5$ nearest neighbors on the right. The pCLE video sequences are represented by their corresponding fused mosaic image. The superimposed visual words, mapped to colors, are highlighting the geometrical structures observed in the pCLE mosaic images. "B" indicates *benign* (i.e. non-neoplastic) and "N" *neoplastic*.

retrieval methods were applied on a database of colonic polyps, but also on a database of Barrett's esophagus as illustrated in Fig. 4.

3 Evaluating pCLE Retrieval Performance

Indirect Retrieval Evaluation based on Classification Due to the inherent difficulty to obtain a ground truth for CBIR on biomedical applications, as pointed out by Müller et al. [6] and by Akgül et al. [7] , it may be advantageous to evaluate retrieval performance in an indirect manner using classification. We consider several pathological classes, for example two classes, and perform k-nearest neighbor classification with leave-one-patient-out (LOPO) cross-validation. This allow us to compare, on the different pCLE databases, the classification performances of our "Dense-Sift" method with respect to several state-of-the-art CBIR methods, namely the statistics-based "Haralick" method [8], the dense "Textons" method [9], the sparse "HH-Sift" method [2], and the "NBNN" classifier [10]. We refer the reader to [4] for a detailed description of these four methods and the evaluation methodology. We demonstrate that, in terms of classification performance, our "Dense-Sift" method significantly outperforms the state-of-the-art methods, on the pCLE database of colonic polyps, as illustrated in Fig. 5, but also on the pCLE database of Barrett's esophagus.

Building a Ground Truth for Perceived Similarity In order to directly evaluate retrieval methods in terms of visual similarity, we aim at construct-

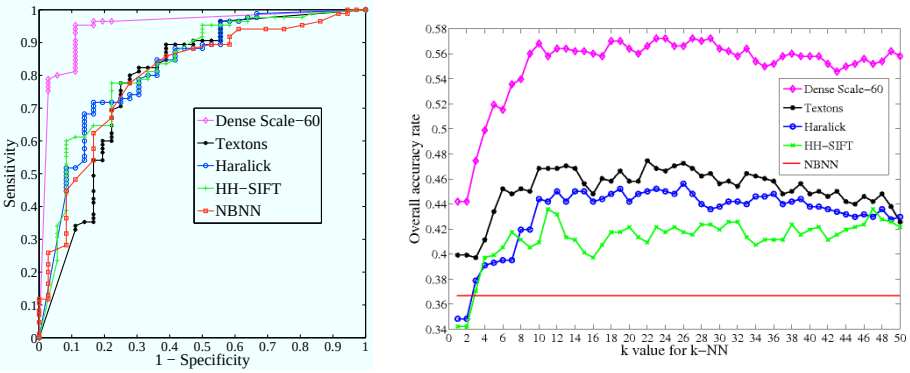

Fig. 5. Left: ROC curves at $k = 5$ neighbors from the LOPO binary classification of pCLE videos of colonic polyps. Right: LOPO 5-class classification of pCLE mosaics by the methods on the colonic polyps database. The NBNN does not depend on k.

ing a ground truth for the perceived similarity between pCLE videos, which is a difficult task because of the subjective appreciation of visual similarities. To facilitate the ground truth generation, we develop in [11] an online survey tool presented in Fig. 6 and available at http://smartatlas.maunakeatech.com, login: MICCAI-User, password: MICCAI2011. This "Visual Similarity Scoring" (VSS) tool allows multiple endoscopists to individually evaluate the visual similarity that they perceived between pCLE videos, according to the four-point Likert scale: "very dissimilar" (L=-2), "rather dissimilar" (L=-1), "rather similar" (L=+1), "very similar" (L=+2). 17 observers, ranging from middle expert to expert in pCLE diagnosis, performed as many scoring processes as they could. In total, 4,836 similarity scores were given for 2,178 distinct video couples. 16.2% of all 13,434 distinct video couples were scored, thus composing a sparse but representative ground truth for perceived similarity.

Direct Retrieval Evaluation against Perceived Similarity From the sparse perceived similarity ground truth obtained on the database of colonic polyps, we are able to perform direct retrieval evaluation by measuring the correlation between the similarity distance based on *visual* signatures and the true perceived similarity. We demonstrate in [11] that, in terms of correlation with the perceived similarity, our "Dense-Sift" method also significantly outperforms the state-of-the-art methods. Furthermore, we define *sparse recall* curves by considering the percentage of videos perceived as "very similar" to the query that were found in the k-neighborhood of the query by the retrieval methods. In Fig. 7 on the left, the *sparse recall* curve of "Dense-Sift" is significantly above the *sparse recall* curves of the state-of-the-art methods, which confirms the previous indirect comparison results.

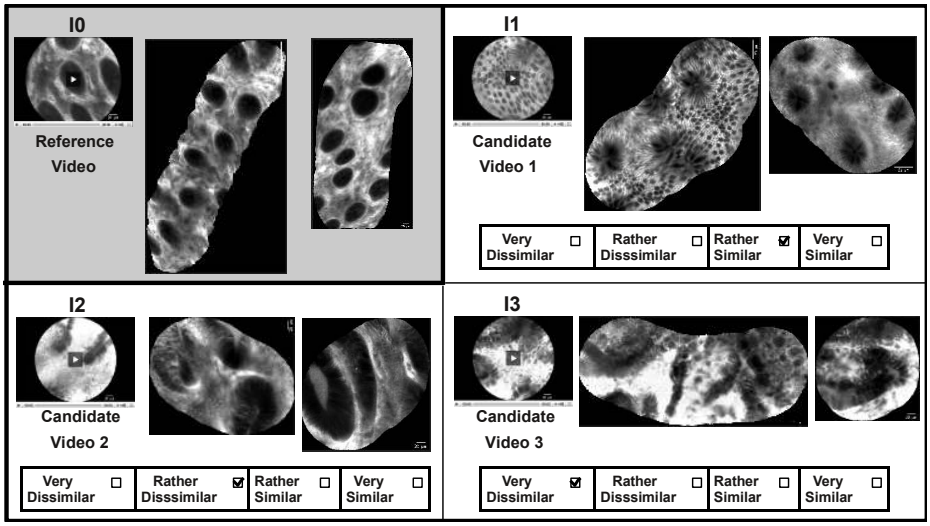

Fig. 6. Schematic outline of the online "Visual Similarity Scoring" tool showing the example of a scoring process, where 3 video couples (I_0, I_1), (I_0, I_2) and (I_0, I_3) are proposed. Each video of the colonic polyps database is summarized by a set of mosaic images, even though the video remains available for viewing.

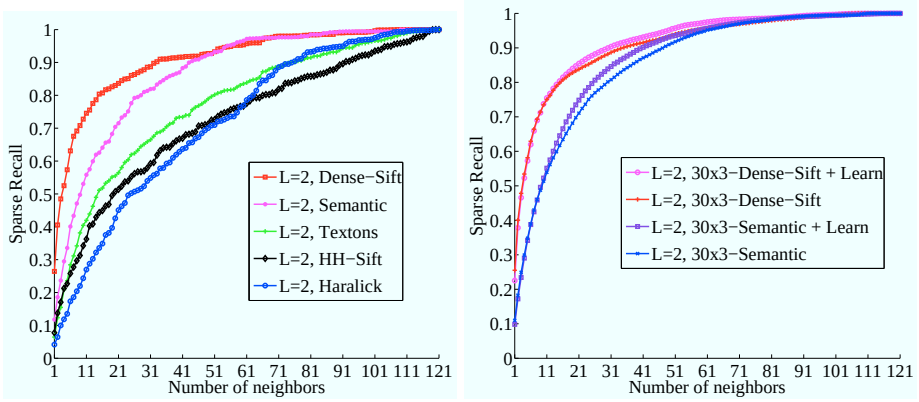

Fig. 7. *Sparse recall* curves on the colonic polyps database, showing the ability of the retrieval methods to capture video pairs perceived as "very similar". Left: without cross-validation, without distance learning. Right: Median of *sparse recall* curves obtained with 30×3-fold cross-validation, before and after distance learning.

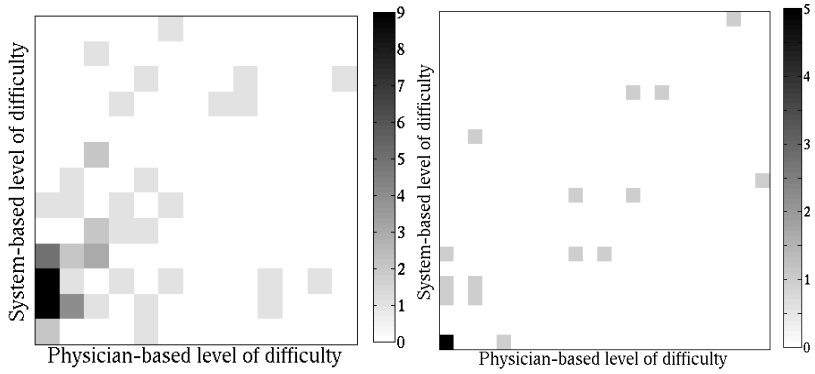

Fig. 8. Joint histograms for the colonic polyps database (left), and for the Barrett's esophagus database (right); x-axis is the *difficulty* experienced by all the physicians and y-axis is our estimated *difficulty*. On the colonic polyps database, 11 physicians, 3 expert and 8 non expert, individually diagnosed 63 videos. On the Barrett's esophagus database, 21 physicians, 9 expert and 12 non expert, individually diagnosed 20 videos.

4 Estimating the Interpretation Difficulty

With the aim of building a self-training simulator for pCLE diagnosis with an adjustable level of difficulty, we propose in [12] to automatically estimate the interpretation *difficulty* associated to a pCLE video by exploiting our retrieval results. For the *difficulty* estimation, we include two main *difficulty* attributes: the first attribute reflects the contextual discrepancies between the video query and its similarity neighborhood, and the second attribute measures the intrinsic ambiguity of the video query with respect to the two pathological classes. Using a robust linear regression model, we leverage the experienced *difficulty* to learn the difficulty predictor from these two attributes. For the learning and the validation steps, another type of ground truth is needed. This ground truth is the *difficulty* experienced by the endoscopists, which is given by the percentage of false pCLE diagnoses, with respect to histology, among several endoscopists. We demonstrate, using permutation tests, that the correlation between the estimated *difficulty* and the experienced *difficulty* is statistically significant on both pCLE databases. Joint histograms can be qualitatively appreciated in Fig. 8. We also notice that correlation coefficients are higher when considering only the non-expert endoscopists.

5 Learning pCLE Semantic and Similarity Distance

Learning pCLE Semantics Aiming at providing the endoscopists with semantic insight into the retrieval results, we investigate pCLE semantic learning. Our semantic ground truth contains 8 binary semantic concepts that are illustrated in Fig. 9 on the left. These semantic concepts were defined by experts in

pCLE in order to support the *in vivo* diagnosis of colonic polyps [13]. Rasiwasia et al. [14] propose a probabilistic method which 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 [15], Kwitt et al. apply 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 [16], we present a Fisher-based approach that leverages the ground-truth data about pCLE semantics in order to compute visual-word-based *semantic* signatures. For each video, our learned *semantic* signature contains 8 floating values, called *semantic weight* $s_j^{j \in \{1,..,8\}}$, such that each *semantic weight* s_j reflects how much the presence of the semantic concept c_j is expressed by the visual words describing the video. Fig. 9 on the right shows the classification performance of the *semantic* signatures with a ROC curve for each semantic concept. The fact that the ROC curves are above the diagonal indicates that each semantic concept contributes to the relevance of the *semantic* signature. Contrary to the "Dense-Sift" *visual* signatures, the *semantic* signatures are not employed for the retrieval task but they are used as an additional information which provides a semantic translation of the visual retrieval outputs. We generate a *sparse recall* curve for the *semantic* signatures as a means to evaluate them from the perceptual point of view and to check their consistency with respect to the *visual* signature. This consistency is shown in Fig. 7 by the fact that the *sparse recall* curve of the *semantic* signature is much closer to the curve of the "Dense-Sift" *visual* signature than the curves of state-of-the-art methods. In Fig. 10, we present a typical retrieval result, where the most similar pCLE videos have been extracted using the *visual* signatures of "Dense-Sift", and where the additional semantic information is provided using a star-plot representation of each visual-word-based *semantic* signature.

Learning Similarity Distance between pCLE Videos In order to boost retrieval performance, we propose in [11, 16] a method to learn an adjusted similarity distance from the perceived similarity ground truth. In a similar way to the method of Philbin et al. [17], our strategy is to shorten the distances between "very similar" videos and to increase the distances between "non very similar" videos. A linear transformation of video signatures is optimized, that minimizes a margin-based cost function differentiating "very similar" video pairs from the others. The *sparse recall* curve associated the transformed *visual* signature is shown in Fig. 7 on the right. We demonstrate that, in terms of correlation with the perceived similarity, the distance learning method allows to improve with statistical significance, the correlation with the perceived similarity. For consistency checking, we also test the distance learning method on the *semantic* signatures, and we show that the transformed *semantics* signature provides a significantly higher correlation with the perceived similarity.

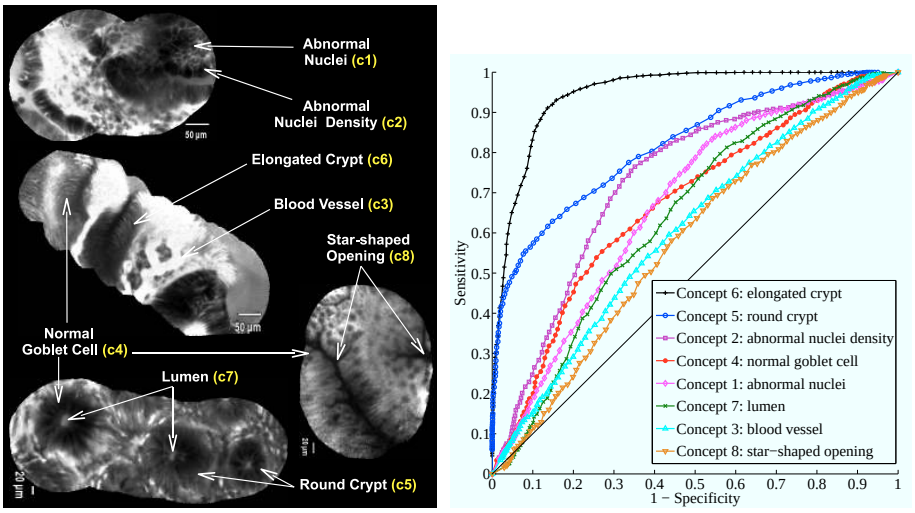

Fig. 9. Left: Examples of training pCLE videos of the colonic polyps database, represented by mosaic images and annotated with the 8 semantic concepts. The two mosaics on the top mosaics show neoplastic colonic polyp, while the two mosaics on the bottom show non-neoplastic colonic polyps. Right: ROC curves showing, for each semantic concept, the classification performance of the visual-word-based *semantic* signature. Each ROC curve associated to a concept c_j is computed with cross-validation by thresholding on the *semantic weight* s_j.

6 Conclusion

In this paper, we presented the first CBIR framework for training and diagnosis support in the field of *in vivo* endomicroscopy. Our main contributions are: 1) a dense Bag-of-Visual-Words method for a pCLE video retrieval that outperforms several state-of-the-art methods; 2) the construction of a perceived similarity ground truth; 3) the development of objective and generic tools for retrieval evaluation; 4) the estimation of pCLE interpretation *difficulty*; 5) a method for pCLE semantics learning to provide the endoscopists with a semantic insight into the retrieval results; 6) a method for perceived similarity learning to boost pCLE retrieval. 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.

Future work will focus on the clinical validation of the "Smart Atlas" and its application to other organs and pathologies. We also plan to investigate spatio-temporal retrieval and to enrich the databases, for example with other metadata to allow for multimodal information retrieval. The "Smart Atlas" tool should allow to increase the diagnostic performance of the endoscopists and to reduce interobserver variability, which should ultimately improve patient outcomes.

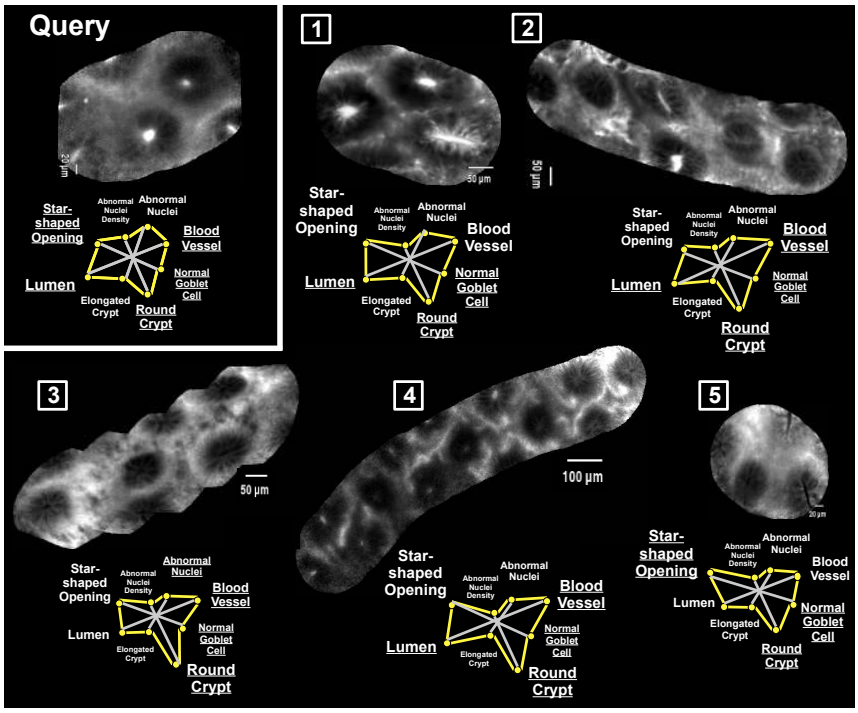

Fig. 10. Typical pCLE retrieval results with semantic extraction on the colonic polyps database, from a non-neoplastic query. The font size of each written semantic concept is proportional to the automatically computed value of the concept coordinate in the star plot. Underlined concepts are those which 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.)

7 Acknowlegments

We would like to thank Pr. Michael B. Wallace and Dr. Anna M. Buchner, who have acquired, analyzed and annotated all the pCLE videos, at the Mayo Clinic in Jacksonville (Florida, US), for their precious contributions.

References

1. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(4) (2009) 591–606
2. Zhang, J., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision **73** (June 2007) 213–238

3. Syeda-Mahmood, T.F., Wang, F., Beymer, D.: Recognition of object categories using affine kernels. In: Multimedia Information Retrieval. (2010) 15–24
4. André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: A smart atlas for endomicroscopy using automated video retrieval. Medical Image Analysis **15**(4) (August 2011) 460–476
5. Vercauteren, T., Perchant, A., Malandain, G., Pennec, X., Ayache, N.: Robust mosaicing with correction of motion distortions and tissue deformation for in vivo fibered microscopy. Medical Image Analysis **10**(5) (October 2006) 673–692
6. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Reisetter, J., Kahn, C.E., Hersh, W.R.: Overview of the clef 2010 medical image retrieval track. In: CLEF (Notebook Papers/LABs/Workshops). (2010)
7. Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-based image retrieval in radiology: Current status and future directions. Journal of Digital Imaging **24**(2) (2011) 208–222
8. Haralick, R.M.: Statistical and structural approaches to texture. In: Proceedings of the IEEE. Volume 67. (1979) 786–804
9. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision **43** (June 2001) 29–44
10. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08). (2008) 1–8
11. André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: 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). (2011) 289–296
12. André, B., Vercauteren, T., Buchner, A.M., Shahid, M.W., Wallace, M.B., Ayache, N.: 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). (2010) 480–487
13. Kiesslich, R., Burg, J., Vieth, M., Gnaendiger, J., Enders, M., Delaney, P., Polglase, A., McLaren, W., Janell, D., Thomas, S., Nafe, B., Galle, P.R., Neurath, M.F.: Confocal laser endoscopy for diagnosing intraepithelial neoplasias and colorectal cancer in vivo. Gastroenterology **127**(3) (2004) 706–13
14. Rasiwasia, N., Moreno, P.J., Vasconcelos, N.: Bridging the gap: Query by semantic example. IEEE Transactions on Multimedia **9**(5) (2007) 923–938
15. Kwitt, R., Rasiwasia, N., Vasconcelos, N., Uhl, A., Häfner, M., Wrba, F.: Learning pit pattern concepts for gastroenterological training. In: Proceedings of the 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'11). (2011) 273–280
16. André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: Learning semantic and visual similarity for endomicroscopy video retrieval. INRIA Technical Report RR-7722, INRIA (August 2011)
17. Philbin, J., Isard, M., Sivic, J., Zisserman, A.: Descriptor learning for efficient retrieval. In: Proceedings of the 11th European Conference on Computer Vision (ECCV'10). (2010) 677–691