Learning Semantic and Visual Similarity for Endomicroscopy Video Retrieval

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

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Learning Semantic and Visual Similarity for Endomicroscopy Video Retrieval

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

N° 7722

Août 2011

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7722--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SOPHIA ANTIPOLIS - MÉDITERRANÉE

Learning Semantic and Visual Similarity for Endomicroscopy Video Retrieval

Barbara André*†, Vercauteren‡, Anna M. Buchner§, Michael B. Wallace¶, Nicholas Ayache‖†

Theme :
Équipes-Projets Asclepios

Rapport de recherche n° 7722 — Août 2011 — 24 pages

Abstract: Traditional Content-Based Image Retrieval (CBIR) systems only deliver visual outputs that are not directly interpretable by the physicians. Our objective is to provide a system for endomicroscopy video retrieval which delivers both visual and semantic outputs that are consistent with each other. In a previous study, we developed an adapted bag-of-visual-words method for endomicroscopy retrieval that computes a visual signature for each video. In this study, we first leverage semantic ground-truth data to transform these visual signatures into semantic signatures that reflect how much the presence of each semantic concept is expressed by the visual words describing the videos. Using cross-validation, we demonstrate that our visual-word-based semantic signatures enable a recall performance which is significantly higher than those of several state-of-the-art methods in CBIR. In a second step, we propose to improve retrieval relevance by learning, from a perceived similarity ground truth, an adjusted similarity distance. Our distance learning method allows to improve, with statistical significance, the correlation with the perceived similarity. Our resulting retrieval system is efficient in providing both visual and semantic information that are correlated with each other and clinically interpretable by the endoscopists.

Key-words: probe-based Confocal Laser Endomicroscopy (pCLE), Content-Based Image Retrieval (CBIR), Bag-of-Visual-Words (BoW) method, semantic learning, similarity distance learning

* Asclepios, INRIA Sophia Antipolis; Mauna Kea Technologies, Paris
† PhD Student in CIFRE thesis
‡ Mauna Kea Technologies, Paris
§ Hospital of the University of Pennsylvania, Philadelphia
¶ Mayo Clinic, Jacksonville, Florida
‖ Asclepios, INRIA Sophia Antipolis

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Apprentissage de la similarité sémantique et visuelle pour la reconnaissance de vidéos endomicroscopiques

Résumé : Les systèmes traditionnels de reconnaissance d'images par le contenu (CBIR) produisent des informations visuelles qui ne sont pas directement interprétables par les médecins. Notre objectif est de concevoir un système pour la reconnaissance de vidéos endomicroscopiques capable de produire des informations visuelles et sémantiques consistantes entre elles. Dans une étude précedente nous avons adapté la méthode des Sac de Mots Visuels pour la reconnaissance en endomicroscopie, et construit une signature visuelle pour chaque vidéo. Dans cette étude, nous commençons par exploiter une vérité terrain contenant des données sémantiques, afin de transformer les signatures visuelles en signatures sémantiques. Celles-ci reflètent dans quelle mesure la présence de chaque concept sémantique est exprimée par les mots visuels qui décrivent les vidéos. A l'aide d'une validation croisée, nous démontrons que nos signatures sémantiques basées mots visuels permettent d'obtenir une performance de reconnaissance supérieure, de manière significative, à celle de plusieurs méthodes de l'état de l'art en CBIR. Dans une deuxième temps, nous proposons d'améliorer les résultats de reconnaissance en apprenant, à partir d'une vérité terrain sur la similarité perçue, une distance de similarité adéquate. La distance apprise est davantage corrélée avec la similarité perçue, de manière significative.

Mots-clés : Endomicrocopie Confocale par Minisondes, reconnaissance d'images par le contenu, méthode des Sac de Mots Visuels, apprentissage d'une sémantique, apprentissage d'une distance de similarité

1 Introduction

The expanding application of Content-Based Image Retrieval (CBIR) methods of computer vision in the medical diagnosis field faces the semantic gap, which was pointed out by Smeulders et al. in [1] and by Akgül et al. in [2], as a critical issue. In CBIR, the semantic gap is the disconnection between the reproducible computational representation of low-level visual features in images and the context-dependent formulation of high-level knowledge, or semantics, to interpret these images. Two medical images being highly similar in appearance may have contradictory semantic annotations. So a CBIR system, which would be only based on visual content, might lead the physician toward a false diagnosis. Conversely, two medical images having exactly the same semantic annotations may look visually dissimilar. So a CBIR system, for which the semantics of the query is unknown, might not retrieve all clinically relevant images. In fact, when interpreting a new image for diagnostic purposes, the physician uses similarity-based reasoning, where *similarity* includes both visual features and semantic concepts. To mimic this process, we aim at capturing the visual content of images using the Bag-of-Visual-Words (BoW) method, and at estimating the expressive power of visual words with respect to multiple semantic concepts. The consistency of the induced visual-word-based *semantic* retrieval could then be tested against perceived similarity ground-truth.

Our medical application is the retrieval of probe-based Confocal Laser Endomicroscopy (pCLE) videos to support the early diagnosis of colonic cancers. pCLE is a recent imaging technology that enables the endoscopist to acquire *in vivo* microscopic video sequences of the epithelium, and thus to establish a diagnosis in real-time. In particular, the *in vivo* diagnosis of colonic polyps using pCLE is still challenging for many endoscopists, because of the high variability in the appearance of pCLE videos and the presence of atypical cases such as serrated adenoma [3]. Examples of mosaic images extracted from pCLE videos are shown in Fig. 1 which also provides an illustration of the semantic gap in endomicroscopy retrieval. In [4] we have developed a dense BoW method, called "Dense-Sift", for the content-based retrieval of pCLE videos. We showed that, when evaluated in terms of pathological classification of pCLE videos, "Dense-Sift" significantly outperforms several state-of-the-art CBIR methods. Parts of this paper are extensions of a preliminary study [5] where we explored pCLE retrieval evaluation and distance learning in terms of perceived visual similarity. Here, our objective is to learn the pCLE similarity distance both in terms of visual appearance and semantic annotations, in order to provide the endoscopists with semantic insight into the retrieval results.

To this purpose, we consider two types of ground-truths presented in Section 2: the first type contains visual similarities perceived by endoscopists between pCLE videos, evaluated on a four-points Likert scale, and the second type contains multiple binary semantic concepts identified by experts in pCLE videos. These eight binary concepts, illustrated in Fig. 2, have been defined to support the *in vivo* pCLE diagnosis of colonic polyps. In Section 3 we shortly described our "Dense-Sift" retrieval method. From the *visual* signatures computed by "Dense-Sift" and from the semantic ground-truth, we build visual-word-based *semantic* signatures using a Fisher-based approach detailed in Section 4. We evaluate the relevance of the resulting *semantic* signatures, first from the semantic point of view, with ROC curves showing classification perfor-

mances for each semantic concept, and then from the perceptual point of view, with *sparse recall* curves showing the ability of the induced retrieval system to capture video pairs perceived as *very similar*. Retrieval performance is also evaluated by measuring the correlation of the induced similarity distance with the perceived similarity ground-truth. In order to improve retrieval relevance, we propose in Section 5 a method to learn an adjusted similarity distance from the perceived similarity ground-truth. A linear transformation of video signatures is optimized, that minimizes a margin-based cost function differentiating *very similar* video pairs from the others. The results shown in Section 6 show that the visual-word-based *semantic* signatures yield a recall performance which is slightly lower than that of the original *visual* signatures computed by "Dense-Sift", but significantly higher than those of several state-of-the-art methods in CBIR. In terms of correlation with the perceived similarity, the retrieval performance of *semantic* signatures is better, with statistical significance, than those of the state-of-the-art methods, and comparable to that of the original *visual* signatures. For both *semantic* signatures and *visual* signatures, the distance learning method allows to improve, with statistical significance, the correlation with the perceived similarity. Our resulting pCLE retrieval system, of which visual and semantic outputs are consistent with each other, should better assist the endoscopist in establishing a pCLE diagnosis.

2 Ground-Truths for Perceived Visual Similarity and Semantic

2.1 pCLE database

Our video database contains 118 pCLE videos of colonic polyps that were acquired from 66 patients for the study of Buchner et al. [6]. The lengths of the acquired pCLE videos range from 1 second to 4 minutes. During the colonoscopy procedures the pCLE miniprobe was in constant contact with the epithelium, so the viewpoint changes between the images of stable pCLE video sequences are mostly in-plane rotations and translations. This is the reason why we can represent any pCLE video as a set of mosaic images built with the video-mosaicing technique of Vercauteren et al. [7], each mosaic image corresponding to a stable subsequence of the video. pCLE mosaic images will not only be used as inputs for our retrieval system, but also as retrieval outputs attached to the extracted similar videos. Indeed, Dabizzi et al. [8] recently showed that pCLE mosaics have the potential to replace pCLE videos for a comparable diagnosis accuracy and a significantly shorter interpretation time.

2.2 Ground-Truth for Perceived Visual Similarity

To generate a pairwise similarity ground-truth between pCLE videos, we designed an online survey tool, called VSS [9], that allows multiple observers, who are fully blinded to the video metadata such as the pCLE diagnosis, to qualitatively estimate the perceived visual similarity degree between videos. The VSS tool proposes, for each video couple, the following four-points Likert scale: *very dissimilar*, *rather dissimilar*, *rather similar* and *very similar*. Because interpreting whole video sequences is time consuming, the VSS supports this task

by making available both the whole video content and for each video, its set of static mosaic images providing a visual summary. Each scoring process, as illustrated in Fig. 3, is characterized by the random drawing of 3 video couples (I_0, I_1), (I_0, I_2) and (I_0, I_3), where the candidate videos I_1, I_2 and I_3 belong to patients that are different from the patient of the reference video I_0, in order to exclude any patient-related biases. 17 observers, ranging from middle expert to expert in pCLE diagnosis, performed as many scoring processes as they could. Our generated ground-truth can be represented as an graph where the nodes are the videos and where each couple of videos may be connected by zero, one or several edges representing the similarity scores. As less than 1% of these video couples were scored by more than 4 distinct observers, it was not relevant to measure inter-observer variability. In total, $4,836$ similarity scores were given for $2,178$ distinct video couples. Thus 16.2% of all $13,434$ distinct video couples were scored. Compared to our preliminary study [5] where 14.5% of all possible video couples were scored, the perceived similarity ground-truth was enriched for this study in order to better differentiate potentially *very similar* video pairs from the others, a goal which is closer to our retrieval purpose.

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

More precisely, in the first mode, the probability of drawing a video couple (I_i, I_j) is proportional to the inverse of the density of $d_{prior}(I_i, I_j)$. In the second mode, the video I_j is one of the 5 nearest neighbors of the video I_i according to the retrieval distance d_{Vis}. A total of $3,801$ similarity scores was recorded with the first mode, and $1,035$ with the second mode.

Although the resulting similarity graph remains very sparse, we will show in Section 6 that it constitutes a valuable ground-truth database for retrieval evaluation and for perceived similarity learning.

2.3 Ground-Truth for Semantic Concepts

All the acquired pCLE videos were manually annotated with $M = 8$ binary semantic concepts describing the observed colonic polyps. These concepts are illustrated on pCLE mosaic images in Fig. 2. In a given pCLE video, each semantic concept is defined as either visible, potentially several times, or not visible at all in the video. The first two concepts, *abnormal nuclei* (c_1) and *abnormal nuclei density* (c_2), which are the most difficult to identify, were annotated by two expert endoscopists. With the support of the modified Mainz criteria identified by Kiesslich et al. [10] six other concepts were annotated: *blood vessel* (c_3), *normal goblet cell* (c_4), *round crypt* (c_5), *elongated crypt* (c_6), *lumen* (c_7) and *star-shaped opening* (c_8). If the semantic j^{th} concept is visible in the video then $c_j = 1$ else $c_j = 0$.

3 From pCLE Videos to Visual Words

Among the state-of-the-art methods in CBIR, the BoW method of Zhang et al. [11], referred to as "HH-Sift", is particularly successful for the retrieval of texture images in computer vision. Whereas "HH-Sift" combines the sparse "Harris-Hessian" detector with the SIFT descriptor, the competitive "Textons" method proposed by Leung and Malik [12] is based on a dense description of local texture features. Adjusting these approaches for pCLE retrieval, we proposed in [4] the "Dense-Sift" method with the following parameters: disk regions of radius 60 pixels, a total of $K = 100$ visual words and dense SIFT description of explicit mosaic images. The image description performed by "Dense-Sift" is invariant to in-plane rotations and in-plane translations changes that are due to the motion of the pCLE miniprobe, and to the affine illumination changes that are due to the leakage of fluorescein used in pCLE. "Dense-Sift" also enables the extension from pCLE image description to pCLE video description by leveraging video mosaicing results. As a result, "Dense-Sift" computes a visual word signature $\mathscr{S}_{\text{Vis}}(I) = (w_1^I, ..., w_K^I)$ for each pCLE video I, where w_k^I is the frequency of the k^{th} visual word in the video I. We define the visual similarity distance $d_{\text{Vis}}(I, J)$ between two videos I and J as the χ^2 pseudo-distance between their visual word signatures computed by "Dense-Sift":

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

Another CBIR method considered as competitive is the "Haralick" [13] method based on global statistical features.

Among the four competitive CBIR methods, "HH-Sift", "Textons", "Dense-Sift" and "Haralick", our "Dense-Sift" method was proved in [4] to be the best method in terms of pathological classification of pCLE videos. "Dense-Sift" will also be proved to be the best method in terms of correlation with the perceived visual similarity, as shown in Section 6. For these reasons, we decided to build the *semantic* signatures of pCLE videos from the *visual* signatures computed by "Dense-Sift".

4 From Visual Words to Semantic Signatures

Among the approaches in bridging the semantic gap, recent methods based on random-walk processes on visual-semantic graphs were proposed by Poblete et al. [14] and by Ma et al. [15]. Latent semantic indexing approaches have also been investigated, for example by Caicedo et al. [16] to improve medical image retrieval. Rasiwasia et al. [17, 18] proposed a probabilistic method which we consider as a reference method for performing a semantic retrieval which is based on visual features. In particular, their approach estimates for each semantic concept the probability that, given a visual feature vector in an image, the semantic concept is present in the image. In [19], Kwitt et al. recently applied this method for learning pit pattern concepts in endoscopic images of colonic polyps. These pit pattern concepts at the macroscopic level can be seen as corresponding to our semantic concepts at the microscopic level. In order to learn semantic concepts from visual words in endomicroscopic videos, we

propose a rather simple method providing satisfactory results. The application of a probabilistic method such as the one in [17] on our data was not successful, certainly because of our relatively small sample size, but we plan to further investigate it. Our proposed method is a Fisher-based approach that estimates the expressive power of each of the K visual words with respect to each of the M semantic concepts.

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

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

where $\mu_p(k,j)$ (resp. $\sigma_p^2(k,j)$) is the mean (resp. the variance) of $\{w_k^I,\ c_j^I = p,\ I \in D^{train}\}$ with $p = 0$ or $p = 1$. We call F the resulting matrix of Fisher's weights. Given a video I of *visual* signature $\mathscr{S}_{\text{Vis}}(I) = (w_1^I, ..., w_K^I)$, we define the *semantic weight* of I with respect to j^{th} semantic concept as the following linear combination: $s_j^I = \sum_{k=1}^{K} F_{k,j} w_k^I$. Thus, the transformation from the *visual* signature $\mathscr{S}_{\text{Vis}}(I)$ into its visual-word-based *semantic* signature $\mathscr{S}_{\text{Sem}}(I) = (s_1^I, ..., s_M^I)$ is given by the equation:

$$\mathscr{S}_{\text{Sem}}(I) = F^T \mathscr{S}_{\text{Vis}}(I) \tag{3}$$

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

$$d_{\text{Sem}}(I,J) = ||\mathscr{S}_{\text{Sem}}(I) - \mathscr{S}_{\text{Sem}}(J)||_{L^2} \tag{4}$$

It thus becomes possible to use our short *semantic* signature of size $M = 8$ in order to retrieve pCLE videos that are the closest to a video query according to the *semantic* distance d_{Sem}. In Section 6 we will demonstrate that, in terms of correlation with the perceived visual similarity, the retrieval performance of the *semantic* distance d_{Sem} is comparable to that of the visual distance d_{Vis}.

In order to provide the endoscopists with a qualitative visualization of *semantic* signatures, we provide an intuitive representation of any *semantic* signature using a star plot of M radii, as shown in Fig. 4. Given a video I and the j^{th} semantic concept, we normalize the *semantic weight* s_j^I into $(s_j^I - \min\{s_j^J, J \in D^{train}\})/(\max\{s_j^J, J \in D^{train}\} - \min\{s_j^J, J \in D^{train}\})$ in order to obtain the coordinate value of I along the j^{th} radius of the star plot. For example, in Fig. 5 the star plots represent, from some tested videos, the visual-word-based *semantic* signatures that have been learned from annotated training videos, such as the ones shown in Fig. 2.

5 Distance Learning from Perceived Similarity

Similarity distance learning has been investigated by rather recent studies to improve classification or recognition methods. Yang et al. [20] proposed a boosted distance metric learning method that projects images into a Hamming space

where each dimension corresponds to the output of a weak classifier. Weinberger and Saul [21] explored convex optimizations to learn a Mahalanobis transformation such that distances between nearby images are shrunk if the images belong to the same class and expanded otherwise. At the level of image descriptors, Philbin et al. [22] have a similar approach that transforms the description vectors into a space where the clustering step more likely assigns matching descriptors to the same visual word and non-matching descriptors to different visual words.

In order to improve the relevance of pCLE retrieval, our objective is to shorten the distances between *very similar* videos and to enlarge the distances between non-*very similar* videos. As the approach of Philbin et al. [22] is closer to our pairwise visual similarity ground-truth, we propose a generic distance learning technique inspired from their method. We aim at finding a linear transformation matrix W which maps given video signatures to new signatures that better discriminate *very similar* video pairs from the other video pairs. We thus consider two groups: D_+ is the set of N_+ training video couples that have been scored with $+2$ and D_- is the set of N_- training video couples that have been scored with $+1$, -1 or -2. We optimize the transformation W by minimizing the following margin-based cost function f:

$$\begin{aligned} f(W,\beta,\gamma) = & \\ & \frac{1}{N+} \sum_{(I,J)\in D_+} L(\beta - d(W\ \mathscr{S}(I), W\ \mathscr{S}(J))) \\ +\gamma\ & \frac{1}{N_-} \sum_{(I,J)\in D_-} L(d(W\ \mathscr{S}(I), W\ \mathscr{S}(J)) - \beta) \end{aligned} \tag{5}$$

where $\mathscr{S}(I)$ is the signature of the video I, $d(.,.)$ is the chosen distance between the video signatures and $L(z) = log(1 + e^{-z})$ is the logistic-loss function. The cost function f has the 3 following parameters: the transformation matrix W, the margin β and the constant parameter γ_s that potentially penalizes either non-*very similar* nearby videos or *very similar* remote videos. We could optimize f with respect to all 3 parameters, but this would make the search for the optimum more sensitive to local minima. We therefore decide to fix the value of β and γ using intuitive heuristics and we are left with the optimization with respect to W alone. As a relevant value for the margin b, we take the threshold on the distances between video signatures that maximizes the classification accuracy between D_+ and D_-. The optimal value of γ_s is then determined using cross-validation. As long as the distance $d(.,.)$ is differentiable, f can be differentiated with respect to W. Given a pCLE video I, its signature $\mathscr{S}(I)$ of size N is mapped to the transformed signature $W^{opt}\ \mathscr{S}(I)$, where W^{opt} is the optimized transformation matrix of size $N \times N$. The learned similarity distance between two pCLE videos I and J is then defined as:

$$d^{learn}(I,J) = d(W^{opt}\mathscr{S}(I), W^{opt}\mathscr{S}(J))) \tag{6}$$

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

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

$$d(W\ \mathscr{S}_{\text{Vis}}(I), W\ \mathscr{S}_{\text{Vis}}(J)) = \chi^2\left(\frac{W\ \mathscr{S}_{\text{Vis}}(I)}{||W\ \mathscr{S}_{\text{Vis}}(I)||_{L^1}}, \frac{W\ \mathscr{S}_{\text{Vis}}(J)}{||W\ \mathscr{S}_{\text{Vis}}(J)||_{L^1}}\right) \quad (7)$$

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

6 Evaluation and Results

6.1 Cross-validation

In order to exclude any learning bias, we used $m \times q$-fold cross-validation, i.e. m random partitions of the database into q subsets. Each of these subsets is successively the testing set and the union of the $q - 1$ others is the training set. To exclude patient-related bias, all videos from the same patient are in the same subset. Given our sparse ground-truth for perceived similarity, q must be not too large in order to have enough similarity scores in each testing set, and not too small to ensure enough similarity scores in the training set. For our experiments, we performed $m = 30$ random partitions of our pCLE video database into $q = 3$ subsets. When computing any performance indicator, we will consider as a robust indicator value the median of all the indicator values computed with cross-validation.

6.2 Evaluation of Semantic Concept Extraction

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

6.3 Retrieval Evaluation Tools

Standard recall curves are a common means of evaluating retrieval performance. However, because of the sparsity of our perceived similarity ground-truth, it is

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

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

Possible indicators of the correlation between the distance computed by a retrieval method and the perceived similarity ground-truth are Pearson correlation π, Spearman ρ and Kendall τ. Compared to Pearson π which measures linear dependence based on the data values, Spearman ρ and Kendall τ are better adapted to the psychometric Likert scale because they measure monotone dependence based on the data ranks [23]. Kendall τ is less commonly used than Spearman ρ but its interpretation in terms of probabilities is more intuitive. To assess statistical significance for the comparison between two correlation coefficients associated to two retrieval methods, we have to perform the adequate statistical test. First, ground-truth data lying on the four-points Likert scale are characterized by a non-normal distribution, so data ranks should be used instead of data values. Second, the rank correlation coefficients measured for two methods are themselves correlated because they both depend on the same ground-truth data. For these reasons, we decide to perform Steiger's Z-tests, as recommended by Meng et al. [24], and we apply it to Kendall τ.

6.4 Results and Discussions

For our experiments, we compared the retrieval performances of "Dense-Sift" with those of "HH-Sift", "Haralick" and "Textons" presented in Section 3 and considered as state-of-the-art method in CBIR. We call "Semantic" the visual-word-based *semantic* retrieval method, "30x3-Semantic" the same method with 30×3 cross-validation and "30x3-Dense-Sift" the "Dense-Sift" with 30×3 cross-validation. "30x3-Semantic+Learn" (resp. "30x3-Semantic+Learn") is the "30x3-Semantic" method (resp. "30x3-Dense-Sift+Learn" method) improved with distance learning.

From the semantic point of view, the performance of the *semantic* signature can be appreciated in the ROC curves shown in Fig. 7. The semantic concepts, from the best classified to the worst classified, are: *elongated crypt*, *round crypt*, *abnormal nuclei density*, *normal goblet cell*, *abnormal nuclei*, *lumen*, *blood vessel* and *star-shaped opening*. The fact that the concept *elongated crypt* is very well classified shows that the visual words clearly express whether this concept is present or not in pCLE videos. As the presence of elongated crypts in a pCLE

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

Table 1: Measures of separability, using Bhattacharyya distance, between the four L-scored histograms H_L shown in Figs 8 and 9 for each retrieval method. For the retrieval methods using 30×3 cross-validation, we computed the median of the Bhattacharyya distances.

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

Table 2: Indicators of correlation between the similarity distance computed by the retrieval methods and the ground-truth. $>$ **M** indicates that the improvement from method **M** is statistically significant, $\sim$ **M** indicates that it is not.

video is a typical criterion of malignancy for the endoscopists, we deduce that *semantic* signatures could be successfully used for pCLE classification between malignant and non-malignant colonic polyps. Although the concepts *blood vessel* and *star-shaped opening* are poorly classified, their contribute to the clinical relevance of the visual-word-based *semantic* retrieval because their ROC curves are above the diagonal.

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

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

In terms of *sparse recall* performances, we observe in Fig. 6 that the retrieval methods from best to worst are: "Dense-Sift+Learn", "Dense-Sift", "Semantic+Learn", "Semantic", "Textons", "HH-Sift" and "Haralick". In particular, perceived similarity distance learning allows to improve slightly recall performance. The fact that "Dense-Sift" outperforms "Semantic" before and after distance learning might be explained by the small size of the *semantic* signatures ($M = 8$) with respect to the larger size of the *visual* signatures ($K = 100$): *semantic* signatures might be too short to discriminate *very similar* video pairs as well as *visual* signatures.

On the superimposed histograms shown in Figs. 8 and 9, we observe qualitatively that "Dense-Sift" and "Semantic" globally better separate the four histograms than "HH-Sift", "Haralick" and "Textons", and that perceived similarity distance learning allows to better separate the histogram H_{+2} from the other histograms. These observations are quantitatively confirmed by the Bhattacharyya distances shown in Table 1. The correlation results shown in Tables 2 and 2 also confirm these findings and demonstrate that, with statistical significance, the similarity distances computed by "Dense-Sift" and "Semantic" are better correlated with the perceived similarity than the similarity distances computed by "HH-Sift", "Haralick" and "Textons". Besides, with statistical significance, the learned similarity distances are better correlated with the perceived similarity than the original distances. These results also show that the correlation performance of "30x3-Semantic+Learn" (resp. "30x3-Semantic") is comparable to that of "30x3-Dense-Sift+Learn" (resp. 30x3-Dense-Sift"), as their difference is not statistically significant.

Looking at the *sparse recall* curves, although the results based on *semantic* signatures are not as good as those based on *visual* signatures, the curve of *semantic* signatures is much closer to the curve of *visual* signatures than the curves of state-of-the-art methods. We can therefore be rather confident in the fact that the *semantic* signatures are informative. *Sparse recall* is only a means to evaluate the relevance of the *semantic* signatures. Indeed, we want to base the retrieval of pCLE videos on visual content and not on semantic annotations, otherwise the retrieval system might retrieve videos that are semantically related but not similar in appearance, in which case the physician might lose trust in the retrieval system. In order to ensure both the higher recall of the visual word retrieval method after distance learning, and the clinical relevance of the semantic information contained in the *semantic* signature, we propose a pCLE retrieval system where the most similar videos are extracted using the "Dense-Sift+Learn" method, and where the star plots representing *semantic* signatures are displayed. Fig. 10 shows some typical results of our pCLE retrieval system with 5 nearest neighbors, with the added semantic ground-truth represented by underlined concepts. In clinical practice, the semantic ground-truth is not known for the video query, but in these retrieval examples it is disclosed for illustration purposes. The extracted pCLE videos, represented as mosaic images, look quite similar in appearance to the query, the first neighbor being more visually similar than the last one. On each star plot, the font size of each written semantic concept is proportional to the normalized value of its *semantic weight*. Semantic concepts written in large characters may or may not be in agreement with the underlined concepts present in the ground-truth. Most importantly, if for a given pCLE video, the semantic ground-truth is very different from the estimated *semantic* signature, then the difficulty to interpret the video for diagnosis purpose might be high, because visual content is not correlated with semantic annotations. Our visual-word-based *semantic* signature would thus have the potential to distinguish ambiguous from non-ambiguous pCLE videos. The remaining disagreements between the learned semantic information and the semantic ground-truth show that, even though we have achieved encouraging results in extracting semantics from visual words, further investigations are still needed to bridge the semantic gap between low-level visual features and high-level clinical knowledge.

7 Conclusion

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

Despite our relatively small pCLE database and despite the sparsity of the perceived similarity ground-truth, our evaluation experiments show that the visual-word-based *semantic* signatures extract, from low-level visual features, a higher-level clinical knowledge which is consistent with respect to perceived similarity. Besides, possible disagreements between the semantic estimation, based on visual features, and the semantic ground-truth could be investigated in order to estimate the interpretation difficulty of pCLE videos, which we explored in a previous study [25] only based on visual words. Future work will focus on more sophisticated methods to learn jointly visual and semantic similarity. Our long-term objective is the clinical evaluation of our visual-semantic retrieval system to see whether it could help the endoscopists in making more accurate pCLE diagnosis.

References

[1] 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.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] C. B. Akgül, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan, and B. Acar, "Content-based image retrieval in radiology: Current status and future directions," *J. Digital Imaging*, vol. 24, no. 2, pp. 208–222, 2011.

[3] O. Khalid, S. Radaideh, O. W. Cummings, M. J. O' Brien, J. R. Goldblum, and D. K. Rex, "Reinterpretation of histology of proximal colon polyps called hyperplastic in 2001," *World J Gastroenterol*, vol. 15, no. 30, pp. 3767–70, 2009.

[4] B. André, T. Vercauteren, M. B. Wallace, A. M. Buchner, and N. Ayache, "Endomicroscopic video retrieval using mosaicing and visual words," in *Proc. ISBI'10*, 2010, pp. 1419–1422.

[5] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "Retrieval evaluation and distance learning from perceived similarity between endomicroscopy videos," in *Proc. MICCAI'11*, vol. to appear, 2011.

[6] 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*, vol. 138, no. 3, pp. 834–842, 2009.

[7] 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.*, vol. 10, no. 5, pp. 673–692, Oct. 2006.

[8] E. Dabizzi, M. W. Shahid, B. Qumseya, M. Othman, and M. B. Wallace, "Comparison between video and mosaics viewing mode of confocal laser endomicroscopy (pCLE) in patients with Barrett's esophagus," *Gastroenterology (DDW 2011)*, 2011.

[9] VSS, "Visual similarity scoring (VSS)," http://smartatlas.maunakeatech.com, login: MICCAI-User, password: MICCAI2011.

[10] R. Kiesslich, J. Burg, M. Vieth, J. Gnaendiger, M. Enders, P. Delaney, A. Polglase, W. McLaren, D. Janell, S. Thomas, B. Nafe, P. R. Galle, and M. F. Neurath, "Confocal laser endoscopy for diagnosing intraepithelial neoplasias and colorectal cancer in vivo," *Gastroenterology*, vol. 127, no. 3, pp. 706–13, 2004.

[11] 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.*, vol. 73, pp. 213–238, Jun. 2007.

[12] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, pp. 29–44, Jun. 2001.

[13] R. M. Haralick, "Statistical and structural approaches to texture," in *Proc. IEEE*, vol. 67, 1979, pp. 786–804.

[14] B. Poblete, B. Bustos, M. Mendoza, and J. M. Barrios, "Visual-semantic graphs: using queries to reduce the semantic gap in web image retrieval," in *Proc. ACM Information and Knowledge Management*, 2010, pp. 1553–1556.

[15] H. Ma, J. Zhu, M. R. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Trans. Multimedia*, vol. 12, pp. 462–473, 2010.

[16] J. C. Caicedo, J. G. Moreno, E. A. Niño, and F. A. González, "Combining visual features and text data for medical image retrieval using latent semantic kernels," in *Proc. Multimedia Information Retrieval*, 2010, pp. 359–366.

[17] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.

[18] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM Multimedia*, 2010, pp. 251–260.

[19] R. Kwitt, N. Rasiwasia, N. Vasconcelos, A. Uhl, M. Häfner, and F. Wrba, "Learning pit pattern concepts for gastroenterological training," in *Proc. MICCAI'11*, vol. to appear, 2011.

[20] L. Yang, R. Jin, L. Mummert, R. Sukthankar, A. Goode, B. Zheng, S. C. H. Hoi, and M. Satyanarayanan, "A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 30–44, 2010.

[21] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.

[22] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor learning for efficient retrieval," in *Proc. ECCV'10*, 2010, pp. 677–691.

[23] V. Barnett, *Sample Survey principles and methods*. Hodder Arnold, 1991.

[24] X.-L. Meng, R. Rosenthal, and D. B. Rubin, "Comparing correlated correlation coefficients," *Psychological Bulletin*, vol. 111, no. 1, pp. 172–175, 1992.

[25] B. André, T. Vercauteren, A. M. Buchner, M. W. Shahid, M. B. Wallace, and N. Ayache, "An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis," in *Proc. MICCAI'10*, no. 6362, 2010, pp. 480–487.

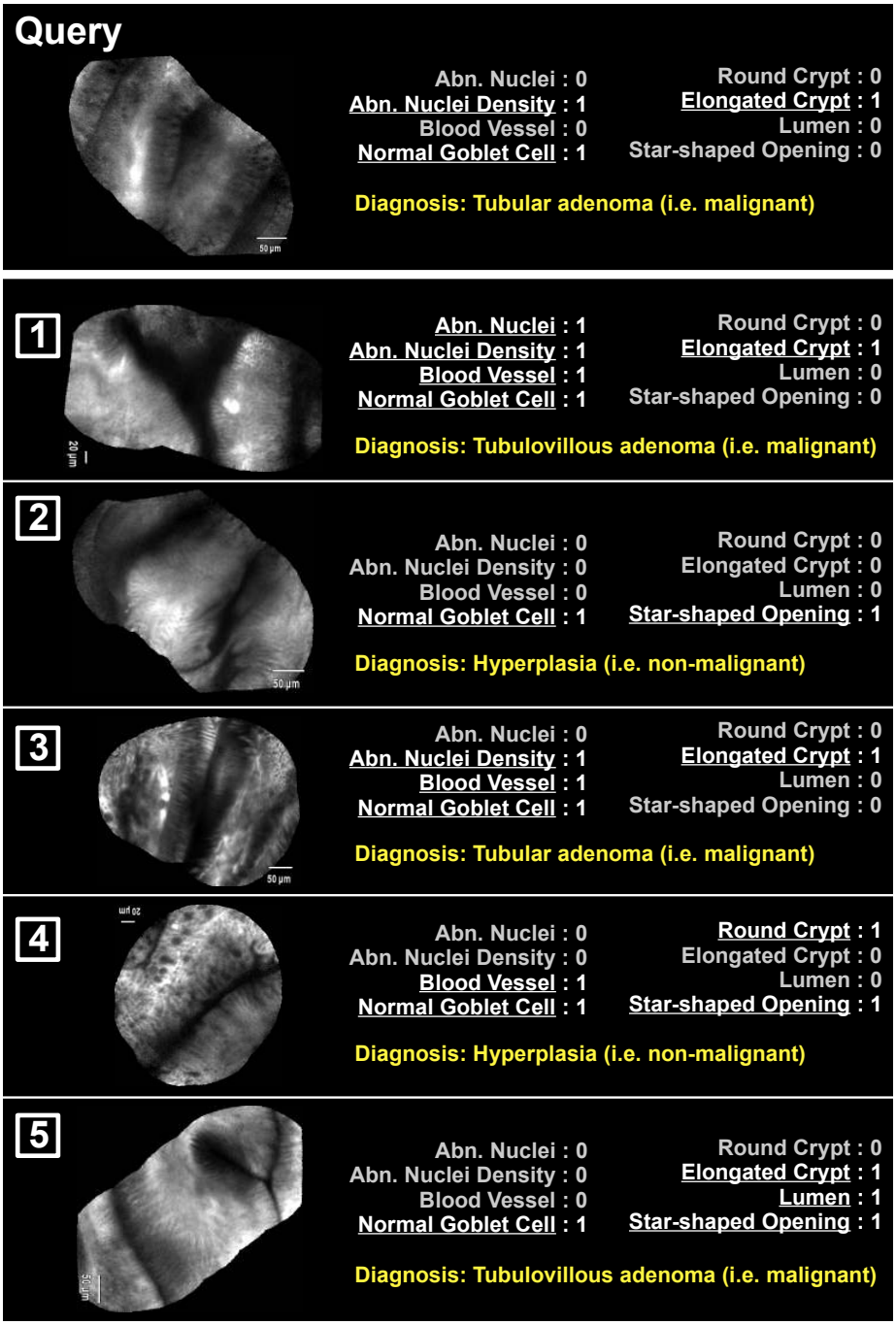


Figure 1: Illustration of the semantic gap: content-based retrieval of visually similar pCLE videos having dissimilar semantic annotations. The 5 most similar pCLE videos are retrieved by the "Dense-Sift" method that only relies on visual features. Semantic concepts which were annotated as present in a given video are underlined. For each video, the pathological diagnosis, either malignant or non-malignant, is indicated below the semantic concepts. For illustration purposes, videos are represented by mosaic images.

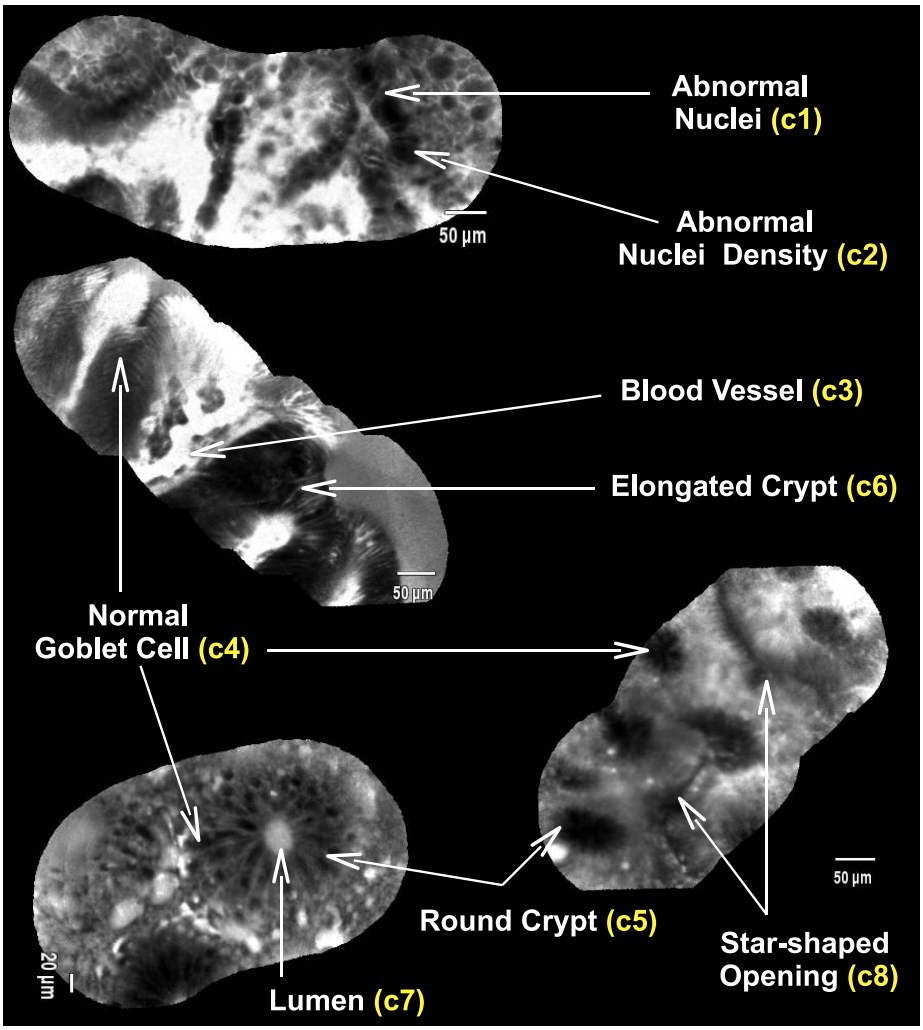


Figure 2: Examples of training pCLE videos represented by mosaic images and annotated with the 8 semantic concepts. The two mosaics on the top mosaics show neoplastic (i.e. malignant) colonic polyp, while the two mosaics on the bottom show non-neoplastic (i.e. non-malignant) colonic polyps.

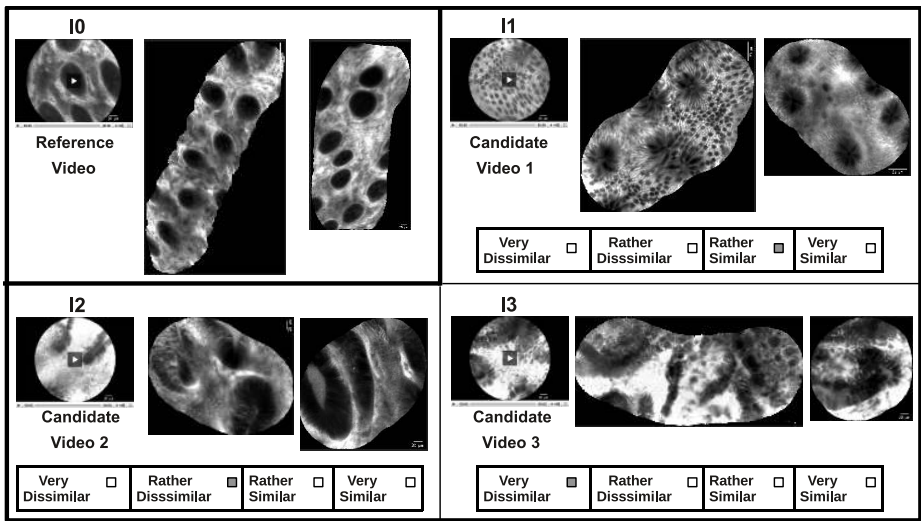


Figure 3: Schematic outline of the online "Visual Similarity Scoring" tool showing the example of a scoring process, where 3 **video couples** (I_0, I_1)**,** (I_0, I_2) **and** (I_0, I_3) **are proposed.** Each video is summarized by a set of mosaic images.

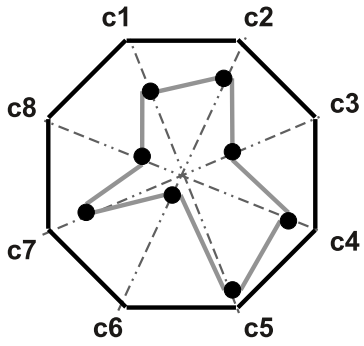


Figure 4: An example of a star plot based on the 8 **semantic concepts.** The coordinate value along the j^{th} radius corresponds to the normalized value of the *semantic* signature at the j^{th} concept.

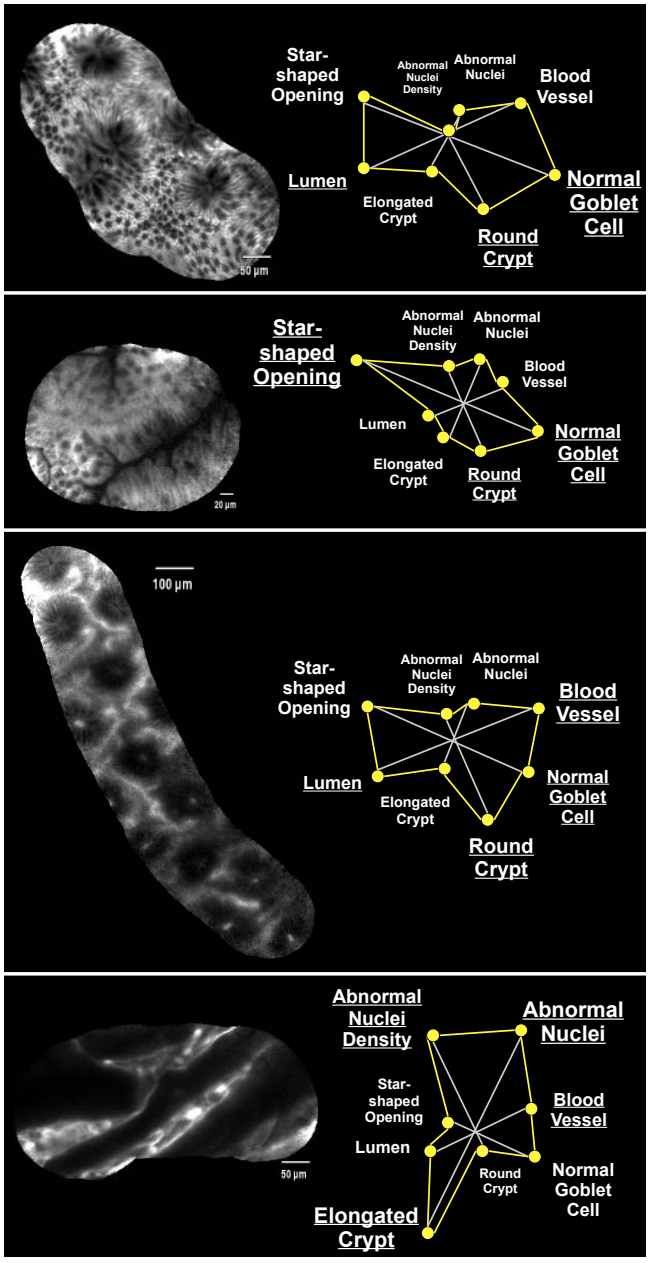


Figure 5: Examples of tested pCLE videos, represented by mosaic images, and visualization of their learned *semantic* signatures using the star plot, as explained in Fig. 4. The font size of each written semantic concept is proportional to the value of the concept coordinate in the star plot. Underlined concepts are those which were annotated as present in the semantic ground-truth. From top to bottom, the first three mosaics show non-neoplastic (i.e. non-malignant) colonic polyps and the fourth mosaic shows a neoplastic (i.e. malignant) colonic polyp.

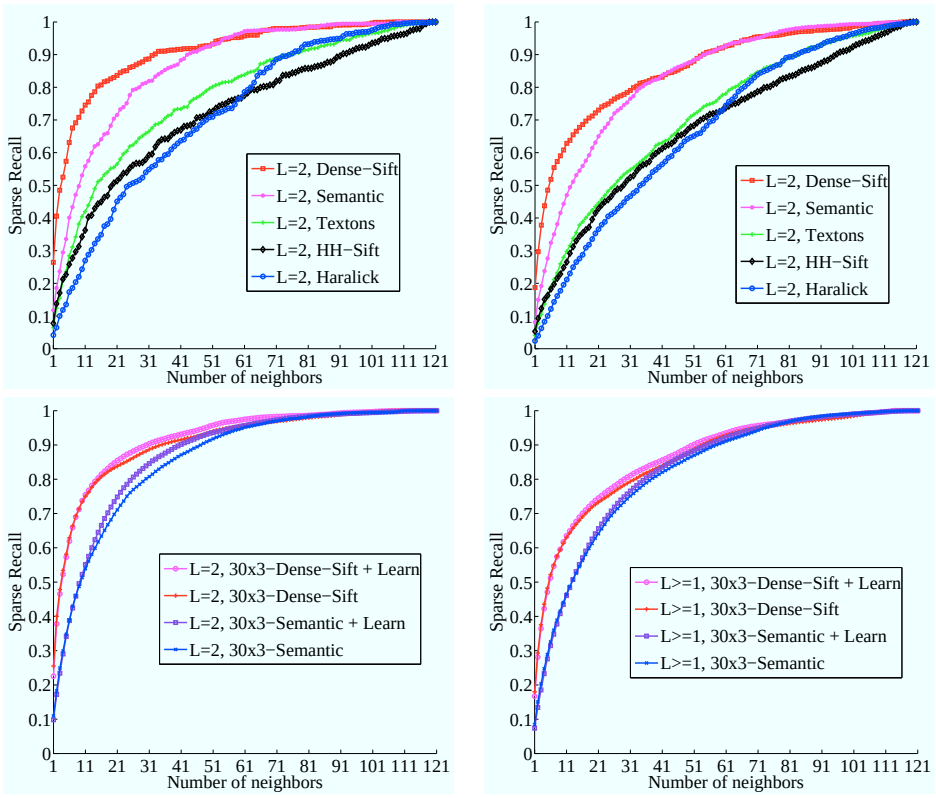


Figure 6: *Sparse recall* curves associated to the retrieval methods in L-scored domains where $L = +2$ (left) or $L >= 1$ (right). For retrieval methods using distance learning, each *sparse recall* curve is the median of the *sparse recall* curves computed with 30×3 cross-validation.

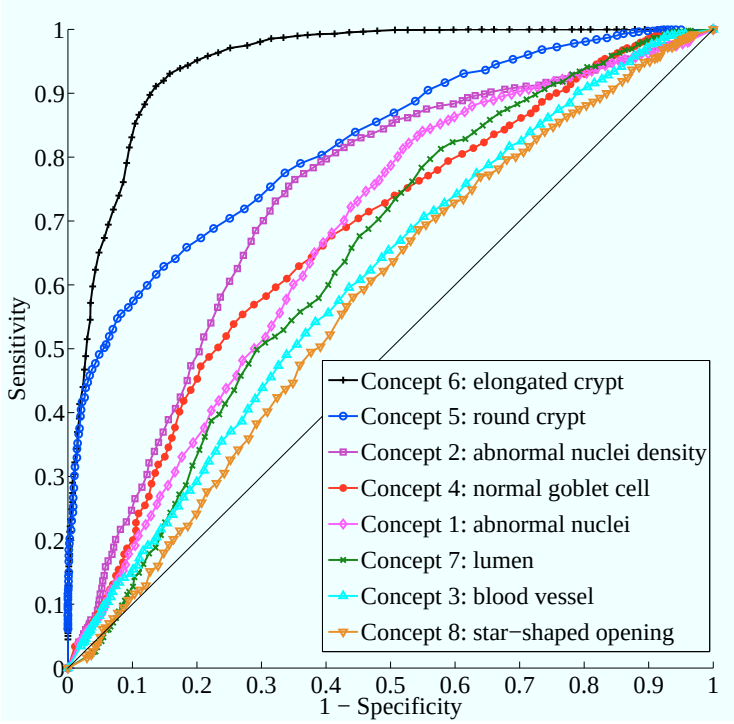


Figure 7: ROC curves showing, for each semantic concept, the classification performance of the *semantic* signature $\mathscr{S}_{\mathbf{Sem}}$. Each ROC curve associated to a concept c_j is the median of the ROC curves computed with 30×3 cross-validation by thresholding on the *semantic weight* s_j.

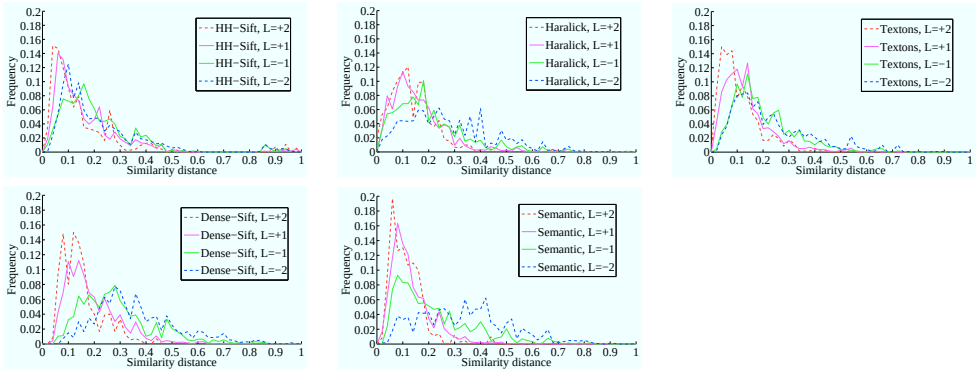


Figure 8: Superimposed histograms H_L of the similarity distances in each L-scored domain. On the top from left to right: "HH-Sift" method, "Haralick" method, "Textons" method. **On the bottom from left to right:** "Dense-Sift" method, "Semantic" method.

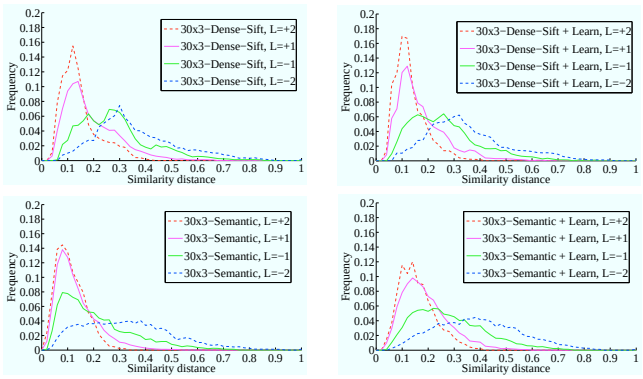
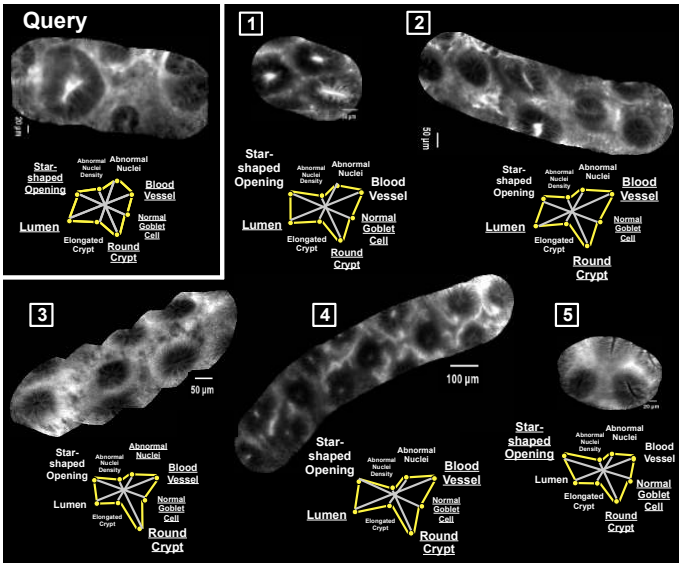


Figure 9: Superimposed histograms H_L of the similarity distances in each L-scored domain. On the top: "30x3-Dense-Sift" method (left) and "30x3-Dense-Sift+Learn" method (right). **On the bottom:** "30x3-Semantic" method (left) and "30x3-Semantic+Learn" method (right). Each histogram is the median of the histograms computed with 30×3 cross-validation.



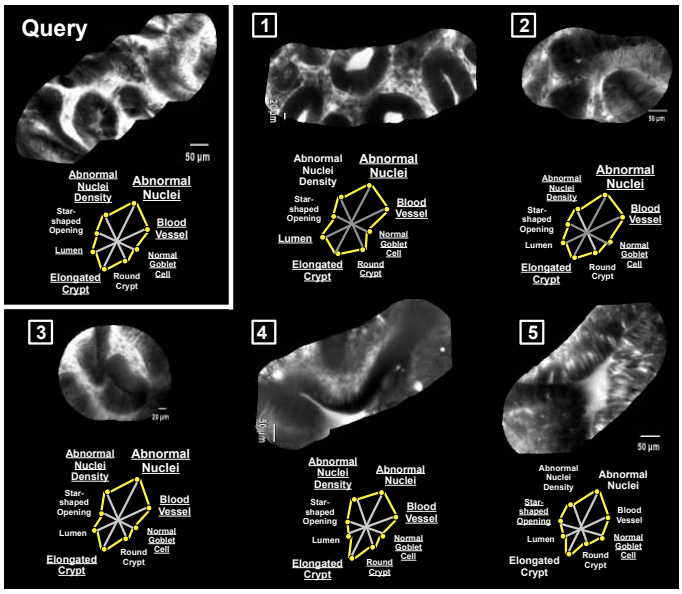


Figure 10: Examples of pCLE retrieval results from a non-neoplastic video query (top) or a neoplastic video query (bottom). The 5 most similar videos are retrieved by "30x3-Dense-Sift+Learn" method. For each video, the star plot representation of its *semantic* signature is provided. The font size of each written semantic concept is proportional to the value of the concept coordinate in the star plot. Underlined concepts are those which were annotated as present in the semantic ground-truth. In practice, the semantic ground-truth is not known for the video query, but it is disclosed here for illustration purposes. For illustration purposes, videos are represented by mosaic images.

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399