Introducing space and time in local feature-based endomicroscopic image retrieval

Barbara André[1,2], Tom Vercauteren[1], Aymeric Perchant[1], Anna M. Buchner[3], Michael B. Wallace[3], and Nicholas Ayache[2]

[1]Mauna Kea Technologies (MKT), Paris, France
[2]INRIA - Asclepios Research Project, Sophia Antipolis, France
[3]Mayo Clinic, Jacksonville, Florida, USA

Abstract. Interpreting endomicroscopic images is still a significant challenge, especially since one single still image may not always contain enough information to make a robust diagnosis. To aid the physicians, we investigated some local feature-based retrieval methods that provide, given a query image, similar annotated images from a database of endomicroscopic images combined with high-level diagnosis represented as textual information. Local feature-based methods may be limited by the small field of view (FOV) of endomicroscopy and the fact that they do not take into account the spatial relationship between the local features, and the time relationship between successive images of the video sequences. To extract discriminative information over the entire image field, our proposed method collects local features in a dense manner instead of using a standard salient region detector. After the retrieval process, we introduce a verification step driven by the textual information in the database and in which spatial relationship between the local features is used. A spatial criterion is built from the co-occurence matrix of local features and used to remove outliers by thresholding on this criterion. To overcome the small FOV problem and take advantage of the video sequence, we propose to combine image retrieval and mosaicing. Mosaicing essentially projects the temporal dimension onto a large field of view image. In this framework, videos, represented by mosaics, and single images can be retrieved with the same tools. With a leave-n-out cross-validation, our results show that taking into account the spatial relationship between local features and the temporal information of endomicroscopic videos by image mosaicing improves the retrieval accuracy.

1 Introduction

With the recent technology of probe-based confocal laser endomicroscpy (pCLE), endoscopists are able to image tissues at microscopic level with a miniprobe, and in real time during ongoing procedure. However, as the acquired pCLE images are relatively new for them, the physicians are still in the process of defining a taxonomy of the pathologies in the images, for instance to differentiate benign tissues and neoplastic, i.e. pathological, tissues of colonic polyps, see Fig. 1 for an illustrative example of such images. To face this clinical challenge, a valuable aid to the physician in establishing a diagnosis would be to provide endomicroscopic images that have a similar appearance to the image of interest and that have been previously diagnosed by expert physicians. Knowing that pathological tissue is characterized by some irregularities in the cellular and vascular architecture, we

aim at retrieving texture information coupled with shape information by using local operators on pCLE images. To serve that purpose, we decided to investigate a modern method for content-based image retrieval (CBIR), the bag-of-visual words (BVW) method [1]. BVW has been successfully used in many applications of computer vision. For example, by applying this method on a large variety of images of natural or artificial textures, the authors of [1] obtained excellent recognition results that are close to 98%.

The standard BVW method detects salient regions in the images and extracts information only on these specific regions. However in pCLE images, the discriminative information is distributed over the entire image field. Contrary to classical methods that apply sparse detectors, we use a dense detector to collect densely the local features in the images. This overcomes the information sparseness problem. Moreover, pCLE images contain characteristic pattern at several scales, in particular the microscopic scale of individual cells and the mesoscopic scale of groups of cells. For this reason, we perform a bi-scale description of the collected image regions. Another problem is that the spatial relationship between the local features is lost in the standard BVW representation of an image, whereas the spatial organization of cells is highly discriminative in pCLE images. So we looked at measuring a statistical representation of this spatial geometry. This was achieved by exploiting the co-occurence matrix of the visual words labeling the local features in the image. After the retrieval process, we introduce the measured spatial criterion in a verification step that allows to remove outliers from the retrieved pCLE images, which are given by the most similar to queried images. Taking into account the spatial relationship between local features is the main contribution of our study, it can be used as a generic tool for many applications of CBIR. Besides, we noticed that the FOV of single still pCLE images may not be large enough for the physicians to see a characteristic global pattern and make a robust diagnosis. As this limitation cannot be solved by the standard methods, we decided to take into account the time information of pCLE video sequences by considering them as objects of interest instead of still images. More precisely, we use image mosaicing [2] to project the temporal dimension of video sequences onto a large FOV image, cf. some resulting mosaics in Fig. 1. With a leave-n-out cross-validation, classification experiments on the pCLE database serve the validation of the methodology: our method outperforms other methods taken as references, by improving the classification accuracy and by providing more relevant training images among the first retrieved images.

2 The bag-of-visual words method

As one of the most popular method for image retrieval, the BVW [1] method aims at extracting a local image description that is both efficient to use and invariant with respect to viewpoint changes, e.g., translations, rotations and scaling, and illumination changes, e.g., affine transformation of intensity. Its methodology consists in first finding and describing local features, then in quantizing them into clusters named visual words, and in representing the image by the histogram of these visual words. The BVW retrieval process can thus be decomposed into

four steps: detection, description, clustering and similarity measuring, possibly followed by a classification step for image categorization.

The detection step extracts salient regions in the image, i.e. regions containing some local discriminative information. In particular, corners and blobs in the image can be detected by the sparse Harris-Hessian (H-H) operator around keypoints with high responses of intensity derivatives. Other sparse detectors like the Intensity-Based Regions (IBR) and the Maximally Stable Extremal Regions (MSER) are also specialized for the extraction of blob features in the images. We refer the interested reader [3] for a survey of these detectors.

Then, each local region can be typically described by the Scale Invariant Feature Transform (SIFT) descriptor. We refer the reader [1] for a survey of this and other powerful descriptors. At the description step, the SIFT descriptor computes, for each salient region, a 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. As a result, the image is represented in a high dimensional space by a set of SIFT description vectors that are invariant by translation, rotation and scale.

To reduce the dimension of the description space, the clustering step, for example based on a standard K-Means, builds K clusters, i.e. K visual words, from the union of the description vector sets gathered from all the N images of the training database. Since each description vector counts for one visual word, an image is represented by a signature of size K which is its histogram of visual words, normalized by the number of its salient regions.

Given these image signatures, it is possible to define a distance between two images as the χ^2 distance [3] between their signature and to retrieve the closest training images as the most similar to the image of interest. The relevance of the similarity results can be quantified by a further classification step, for instance based on a standard nearest neighbors procedure that weights the votes of the k-nearest neighbors by the inverse of their χ^2 distance to the signature of the queried image, so that the closest images are the most determinant. Besides, performing image classification is a way to validate a new retrieval method by comparing it with other methods.

3 Including spatial and temporal information

When we applied the standard BVW method on pCLE images, we obtained rather poor classification results, as presented in Section 4, and the presence of many retrieval outliers. To improve the accuracy of endomicroscopic image retrieval, we decided to include both spatial and temporal information contained in the pCLE images. By locally testing on pCLE images the numerous sparse detectors listed by [3], we first observed that a large number of salient regions sparsely extracted by these standard detectors do not persist between two highly correlated successive images taken from the same video. To overcome the persistence problem and take into account all the information in the images, we use a dense detector contrarily to the standard method. This is consistent with the fact that local information appears to be densely distributed over the entire field

of the pCLE image. The dense detector is made of overlapping disks localized on a dense regular grid, such that each disk covers a possible image pattern at microscopic level.

We also noticed that the endoscopists establish their diagnosis on pCLE images from the regularity of the cellular architecture in the colonic tissue [4], where goblet cells and crypts are both round-shaped characteristic patterns, but where a crypt has larger size than its surrounding goblet cells so it must not be recognized as the same object. In order to be sensitive to scale changes, our method looked at describing local disk regions at various scales that are not automatically computed, for example by choosing a microscopic scale for individual cell patterns and a mesoscopic scale for larger groups of cells. This leads us to represent an image by several sets of description vectors that are scale-dependent, resulting in several signatures for the image that are then concatenated into one larger signature.

This previous observation also suggests that the spatial organization of the goblet cells must be included in the retrieval process because it is substantial to differentiate benign tissues from neoplastic tissues. The authors of [5] previously proposed adding a geometrical verification to take 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. In like manner, our idea is to introduce a geometrical verification process after the retrieval process, but based on the assumption that the spatial relationships between the local features are only statistically the same in the images with similar appearance. To introduce spatial information, we took advantage of the dense property to define the adjacency between two visual words as the 8-adjacency between the two disk regions that are labeled by them on the detection grid. Thus, we are able to store in a co-occurence matrix M of size $K \times K$ the probability for each pair of visual words of being adjacent to each other. In order to best differentiate the images of the benign class from the images of the pathological class, we looked at the most discriminative linear combination W of some elements m of M. This is achieved by a linear discriminant analysis (LDA) which uses the textual diagnostic information in the database. The LDA weights are given by $W = \Sigma^{-1} (\mu_1 - \mu_2)$, where Σ is the covariance matrix of the elements m of M in all training images and μ_i is the mean of the elements m of M in the training image belonging to the class i. From these weights W, we computed the spatial criterion $\alpha = Wm$ for each retrieved image and compared it with the α value of the image of interest. By thresholding the α value during a verification process, outliers are rejected and the first retrieved training images are more relevant.

Expert physicians pointed out that some characteristic global patterns are too partially visible on single still pCLE images to make a robust diagnosis: two still images may have a very similar appearance but be attached to contradictory diagnoses. To address this problem, the time dimension of pCLE videos needs to be exploited, by including in the retrieval process the temporal relationship between successive images from the same video sequence. The study reported

by [6] proposes a method for video retrieval using the spatial layout of the image regions, but this method has been designed for object matching, which is not our objective. Since successive frames from pCLE videos are only related by viewpoint changes, our approach uses the image mosaicing of [2] to project the temporal dimension of a video sequence onto one image with a larger FOV and of higher resolution. Thus, mosaics can be queried and retrieved in the same way as still images.

4 Experiments and discussion

At the *anonymized for review*, the Cellvizio® system, MKT, Paris, was used to image colonic polyps during surveillance colonoscopies in 54 patients. On each acquired video sequence, expert physicians established a pCLE diagnosis [4] that differentiates pathological sequences from benign ones. The video sequences contain from 5 to over a thousand frames and each frame is an image of diameter 500 pixels corresponding to a FOV of 240 µm. To build our pCLE database, we considered a subset of these sequences by discarding those whose quality was insufficient to perform a reliable diagnosis. In each of the remaining 52 video sequences, we selected groups of successive frames according to the length of the sequence. The resulting database is composed of $N = 1036$ still pCLE images and $N' = 66$ pCLE mosaics, half of the data coming from benign sequences and half from pathological ones.

In pCLE images, the disk regions containing information at mesoscopic scale have a radius value $\rho_1 = 40$, while the radius value of those containing information at microscopic scale is $\rho_2 = 15$. For the dense detector, we then chose $\delta = 20$ pixels of grid spacing in order to get a reasonable overlap between adjacent regions. Among the values from 30 to 1500 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. To prevent overfitting, as the size of our pCLE database is still rather small, especially concerning the number of mosaics, the number of LDA weights in the computation of the spatial criterion α had to be restricted. For the elements m of the co-occurence matrix M, we only considered the K diagonal elements of the matrix M build from the visual words of large radius 40, observing that the overlapping regions of radius 40 have a sufficient spatial correlation, better than those of radius 15. The good values of the threshold θ_α were chosen by analysing the distribution of α across the benign and pathological images: 2 when retrieving still images from a queried still image and 0.5 when retrieving still images from a queried mosaic.

The classification results of our method are presented in Fig. 2 and compared with the following methods taken as references: the standard sparse scale invariant SIFT method, the statistical approach of Haralick features [7, 8] and the texture retrieval method of Textons [9]. To ensure a non-biased classification, our validation scheme retrieves k nearest images in the training set with training images not belonging to the video sequence of the image being queried, i.e. a leave-n-out cross-validation where n is the number of frames in the video of the queried image. According to the accuracy, sensitivity and specificity rates yielded by each method on the still images of the pCLE database, our retrieval

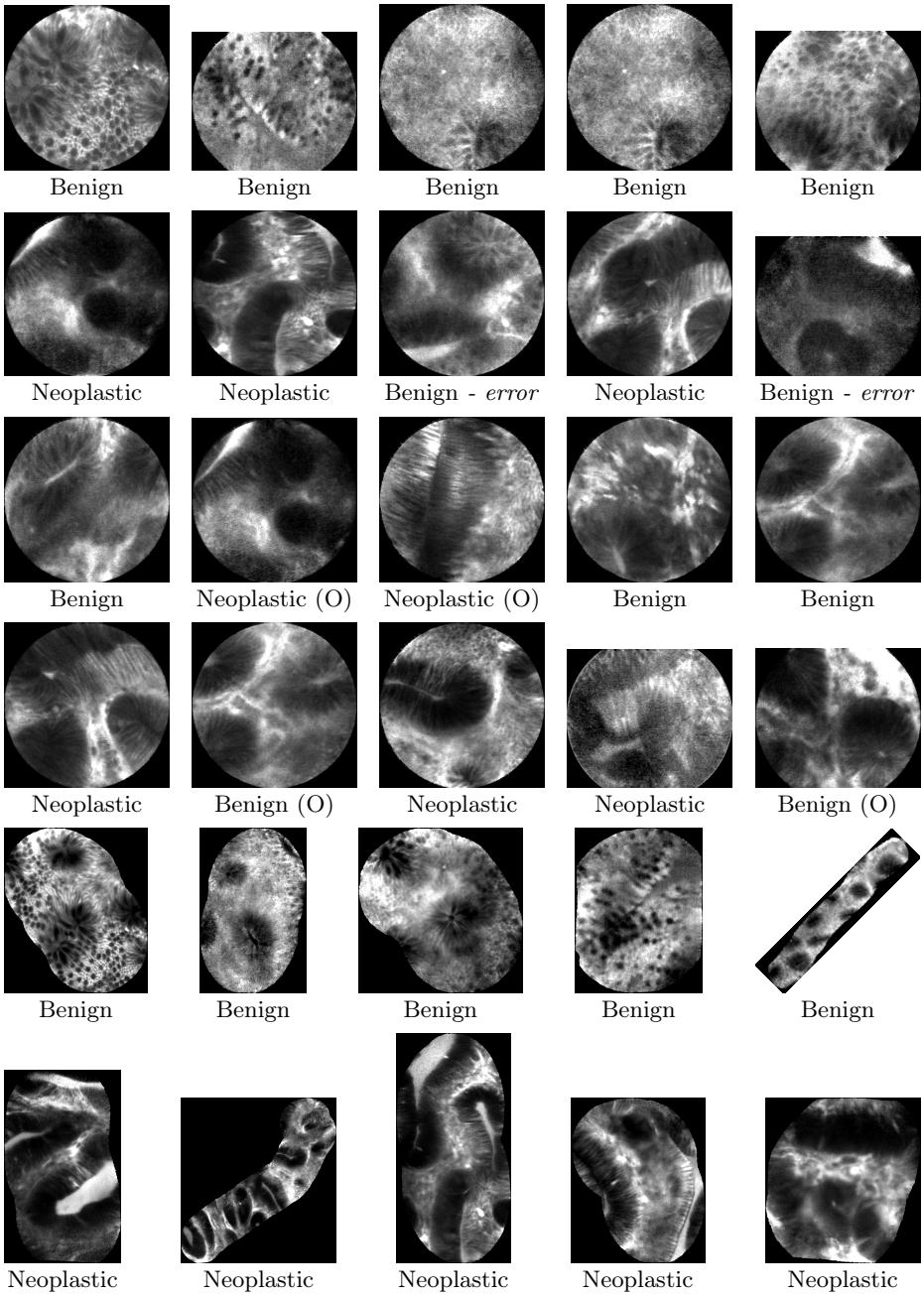
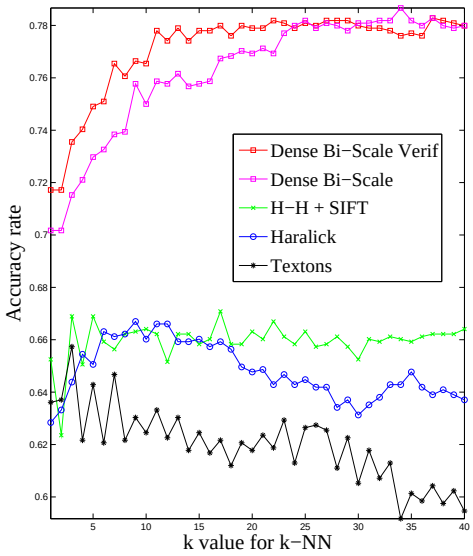


Fig. 1: Six rows of similar pCLE images or mosaics provided by our retrieval method. From left to right on each row: the queried image, and its first, second, third and fourth most similar images. An outlier rejected by the spatial verification process is indicated by the letter **(0)**, and is an *error* otherwise. FOV of the images: 240 µm. FOV of the mosaics: from 260 µm to 1300 µm.



Method / kNN	k=1	k=4	k=10
(I, I) DBV			
Acc.	**71.7**	**74.0**	**76.5**
Sens.	71.8	74.9	77.7
Spec.	71.6	73.1	75.3
(I, I) DB			
Acc.	**70.2**	**72.1**	**75.0**
Sens.	69.8	71.1	73.8
Spec.	70.6	73.3	76.3
(I, I) HH SIFT			
Acc.	**65.3**	**65.1**	**66.4**
Sens.	76.6	75.8	76.6
Spec.	52.7	53.1	55.1
(M, M) DB			
Acc.	**86**	**89**	**82**
Sens.	88	88	78
Spec.	85	91	85
(M, I) DBV			
Acc.	**71.2**	**81.8**	**83.3**
Sens.	96.9	100.0	100.0
Spec.	47.1	64.7	67.7
(M, I) DB			
Acc.	**71.2**	**80.0**	**78.8**
Sens.	90.6	93.8	90.6
Spec.	52.9	67.7	67.7

Fig. 2: Left: Classification accuracies with leave-n-out cross-validation. Right: Results for **k** nearest neighbors, where **M** means **Mosaic** and **I** means **Image** in the configuration (***Queried***, ***Retrieved***). Our proposed method is referred to as Dense Bi-Scale Verif (**DBV**) if it includes spatial verification and Dense Bi-Scale (**DB**) otherwise.

method including spatial information is the most efficient, with an accuracy rate of 78.2% for $k = 22$ neighbors, which is 11.5 points better than the standard SIFT method. The gain of accuracy can be decomposed in 10.2 points for the choice of a dense detector and a bi-scale SIFT description, and 1.3 points for the verification process on the spatial criterion. It is also worth mentioning that with the spatial verification, fewer nearest neighbors are necessary to classify the query at a given accuracy. For $k = 4$ neighbors, some illustrative examples of the image retrieval results are shown in Fig. 1, where the outliers that have been rejected by the verification process are indicated by the letter O.

Moreover, when including both spatial and temporal information by querying mosaics, our classification results are much better. Since mosaics contain more information than single images, their content-based neighborhood is more representative of their pathological neighborhood, so they can be better classified by a smaller number k' of nearest neighbors. Indeed, if we retrieve still images for queried mosaics, the classification accuracy is 83.3% for $k' = 10$ neighbors, which demonstrates the robustness of our retrieval method applied on heterogeneous data with different resolution. For the retrieval of still images from queried mosaics, the poor specificity can be explained by the fact that a mosaic annotated as neoplastic may contain some benign patterns which induce the retrieval of single benign images and classify it as benign. However the expert physicians diagnose a pCLE video sequence as neoplastic as soon as it contains neoplastic patterns, even when some benign tissue is imaged. Besides, if we retrieve mosaics for queried mosaics, the classification accuracy is 89%. Thus, even though we only have a small number of mosaics, including time dimension in mosaics provides us proof of concept results for endomiscroscopic video retrieval. For the

retrieval of mosaics from queried mosaics, including the spatial information does not improve the classification results because of the overfitting phenomenon: indeed, the number of LDA weights, 100, is bigger than the total number of mosaics, $N' = 66$.

5 Conclusion

Using visual similarity between a given image and medically interpreted images allowed us to provide the physicians with semantic similarity, and thus could potentially support their diagnostic decision. Although our experiments are focused on a relatively small training dataset, the classification results constitute a validation of our generic methodology. By taking into account the spatio-temporal relationship between the local feature descriptors, the first retrieved endomicroscopic images are much more relevant. 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-occurence matrix of visual words at several scales. Besides, the learning step of the retrieval process could leverage the textual information of the database and incorporate the spatial information of multi-scale co-occurence matrices into descriptors. As for introducing 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 to work on more accurate visual words and better combine spatial and temporal information.

References

1. Zhang, J., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vis. **73** (June 2007) 213–238
2. 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. Med. Image Anal. **10**(5) (October 2006) 673–692
3. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. Int. J. Comput. Vis. **65** (November 2005) 43–72
4. Buchner, A., Ghabril, M., Krishna, M., Wolfsen, H., Wallace, M.: High-resolution confocal endomicroscopy probe system for in vivo diagnosis of colorectal neoplasia. Gastroenterology **135**(1) (July 2008) 295
5. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. Volume I. (October 2008) 304–317
6. Sivic, J., Zisserman, A.: Efficient visual search for objects in videos. Proc. IEEE **96** (April 2008) 548–566
7. Haralick, R.: Statistical and structural approaches to texture. In: Proc. IEEE. Volume 67. (1979) 786–804
8. Srivastava, S., Rodriguez, J., Rouse, A., Brewer, M., Gmitro, A.: Computer-aided identification of ovarian cancer in confocal microendoscope images. J. Biomed. Opt. **13**(2) (March/April 2008) 024021
9. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. Int. J. Comput. Vis. **43** (June 2001) 29–44