ENDOMICROSCOPIC IMAGE RETRIEVAL AND CLASSIFICATION USING INVARIANT VISUAL FEATURES

B. André[1,2], T. Vercauteren[1], A. Perchant[1], A. M. Buchner[3], M. B. Wallace[3], N. Ayache[2]

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

ABSTRACT

This paper investigates the use of modern content based image retrieval methods to classify endomicroscopic images into two categories: neoplastic (pathological) and benign. We describe first the method that maps an image into a visual feature signature which is a numerical vector invariant with respect to some particular classes of geometric and intensity transformations. Then we explain how these signatures are used to retrieve from a database the k closest images to a new image. The classification is finally achieved through a procedure of votes weighted by a proximity criterion (weighted k-nearest neighbors). Compared with several previously published alternatives whose maximal accuracy rate is almost 67% on the database, our approach yields an accuracy of 80% and offers promising perspectives.

***Index Terms*—** Endomicroscopy, content-based image retrieval, Bag of Visual Words (BVW) method, k-nearest neighbors classification

1. INTRODUCTION AND AIMS

Probe-based confocal laser endomicroscopy (pCLE) is a new technology which enables dynamic microscopic imaging of tissues *in vivo* with a miniprobe during ongoing endoscopy. Our study investigates the application of image retrieval and classification methods to pCLE images of colonic polyps, which would aid the physician in differentiating benign tissues and neoplastic (pathological) tissues. Due to the very specific nature of endomicroscopic images (see Fig. 1 and Fig. 2), where membranes and nuclei are not always clearly visible, typical criteria used by classical computer aided diagnosis methods (e.g. nucleocytoplasmic ratio in histological slices) cannot be applied. Moreover, the taxonomy of pathologies in endomicroscopic images is still under active construction by the physicians, who are discovering their complex underlying semantic. To face this difficulty, we investigated some methods successfully applied in the field of computer vision, where important progress has been recently achieved using local operators for invariant image description and classification. Indeed, the authors of [1] reported excellent recognition results on textured images with a Bag of Visual Words [2] (BVW) method: the method reaches classification results close to 98% on a large variety of images of natural or artificial textures at various scales. This motivated us to adapt their approach to our pCLE images which also tend to contain discriminative texture information (coupled with shape information) at various scales. Thus, our objective is to find the images which are the most similar to a given image by exploring a content based image retrieval (CBIR) approach, and to quantify the relevance of the similarity results by performing a supervised binary classification of the database.

Section 3 explains the contributions of our methodology: the use of a dense detector for salient regions to describe the information over the entire image field, the concatenation of signatures of the same image that is described at different scales (each scale corresponding to a physical group of image patterns), and the supervised selection of the most discriminative visual words to improve the classification results. The performance comparisons between several methods applied to the training set of pCLE images is presented in Section 4; in particular we show that, with a leave-n-out cross-validation, our method outperforms the method of [1] mentioned above.

2. MATERIALS

For our study, colonic polyps were imaged at the Mayo Clinic in Jacksonville using the Cellvizio® system (MKT, Paris), with a prior administration of fluorescein, during surveillance colonoscopies in 54 patients. On each acquired video sequence, the expert physicians established a pCLE diagnosis [3] differentiating pathological sequences from benign ones, according to the presence or not of neoplastic tissue which is characterized by some irregularities in the cellular and vascular architectures. The video sequences contain from 5 to over a thousand frames (images with a circular shape of diameter 500 pixels and with a field-of-view of 240 μm). We considered a subset of these sequences by discarding those whose quality was insufficient to perform a reliable diagnosis, or whose pCLE diagnosis was not the same as the "gold standard" (the "gold standard" for the polyp was established by a pathologist, after histological review). In each of the 52 video sequences that were retained, we selected groups of successive frames according to the length of the sequence.

The resulting database is composed of $N = 1036$ endomicroscopic images, half of the data coming from benign sequences and half from pathological ones.

3. METHODS

The primary goal of local methods for object recognition is to ensure extraction and description of features invariant w.r.t. viewpoint changes (e.g., translations, rotations and scaling) and illumination changes (e.g., affine transformation of intensity). One of the most popular methods for image retrieval using invariant features is the BVW [1, 2] method, that couples for example the sparse Harris-Hessian (H-H) detector of salient regions with the Scale Invariant Feature Transform (SIFT) [4] descriptor of these regions. Given the success of this method, we decided to develop our methodology by adapting the BVW to pCLE images. Our approach is composed of three steps: the detection step, the description step and the classification step.

1) The detection step consists in selecting salient regions in an image, i.e. regions containing some local discriminative information. It is worth noticing that the physicians establish their diagnosis from the regularity of the cellular architecture in the colonic tissue, where goblet cells and crypts are round-shaped or tubular-shaped. For this reason, we looked at extracting blob features in the images by some sparse detectors, like the H-H, the Intensity-Based Regions (IBR) and the Maximally Stable Extremal Regions (MSER) detectors (see for example [5] for a survey of these methods). In particular, the H-H operator detects corners and blobs in the image around key-points with high responses of intensity derivatives for at least two distinct gradient directions. In our pCLE images, we observed that a large number of salient regions extracted by these sparse detectors do not persist between two highly correlated successive frames, although some of them were localized in the center of crypts or in the neighborhood of other colonic patterns. The weak robustness of sparse detectors applied to pCLE images explains the "poor" classification results presented in Section 4; it could be detailed in another study. To take into account local information over the entire image field, the idea is to use a dense detector made of overlapping disks of constant radius distributed on a dense regular grid, such that each disk covers a possible image pattern at microscopic level.

2) For the description step, the standard SIFT descriptor computes a 128-bin gradient histogram as a description vector for each salient region centered on the key-points at optimal scales, the gradient orientations being normalized with respect to the main orientation of the salient region. As a result, each image is represented by a set of description vectors in a high dimensional space. To reduce the dimension of the description space, a standard K-Means clustering builds K clusters named "visual words" from the union of the vector sets representing all the N images of the database. Thus, an image is represented by a signature of size K which is its normalized histogram of visual words, each description vector counting for one visual word. The standard BVW method is at the same time translation, rotation and scale invariant. However, the level of required invariance depends on the problem at hand: we desire invariance by translation and rotation, but we do not want invariance by scaling because in colonic tissue, round-shaped crypts have larger size than round-shaped goblet cells and thus must not be recognized as the same object. For the feature description to be sensitive to scale changes, we decided to perform multiple SIFT descriptions, each of them being associated to a different value of disk radius: for instance, two different radius values lead to represent an image by two sets of description vectors, hence by two signatures that are then concatenated into one larger signature.

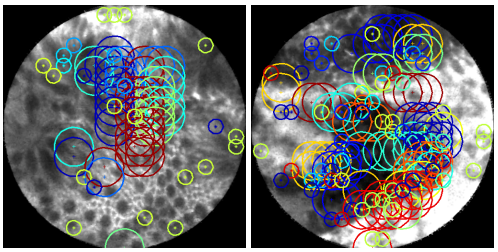

Fig. 1: Disk regions (of radii 15 and 40 pixels) associated to the most discriminative visual words (represented by colors). Left: Benign pCLE image. Right: Neoplastic (pathological) pCLE image. Field of view of the images: 240 μm.

3) The classification step is a standard nearest neighbors procedure: given these image signatures, it is now possible to define a distance between two images as the χ^2 distance between their signature. Besides, we weighted the votes of the k-nearest neighbors by the inverse of their χ^2 distance to the tested image signature, so that the closest images are the most determinant. We also improved the classification performance by selecting in a supervised way the most discriminative visual words, i.e. those minimizing the intra-class distances while maximizing the inter-class distances. For each class C of images, we considered the distribution $p(w|C)$ of the number of occurrences of a visual word w in the images belonging to the class C. The discriminative power $f(w)$ of the visual word w is chosen by using the Fisher criterion which can be expressed as the Mahalanobis distance between the two distributions $p(w|\text{Benign})$ and $p(w|\text{Pathological})$: $f(w) = (\mu_1 - \mu_2)^2/(0.5\,(\sigma_1^2 + \sigma_2^2))$, where μ_i and σ_i are respectively the mean and the variance of the distribution of w in the images belonging to class i. Furthermore, by reducing the number of visual words, the size of image signatures is decreased, so the image retrieval and classification processes run faster.

4. RESULTS

For the dense detector, a large disk radius of $\rho_1 = 40$ pixels is relevant to cover groups of cells with a disk; a smaller disk of radius $\rho_2 = 15$ pixels allows to cover at least one cell in the images (see Fig. 1). Given these radii values, we chose $\delta = 20$ pixels of grid spacing in order to get a rea-

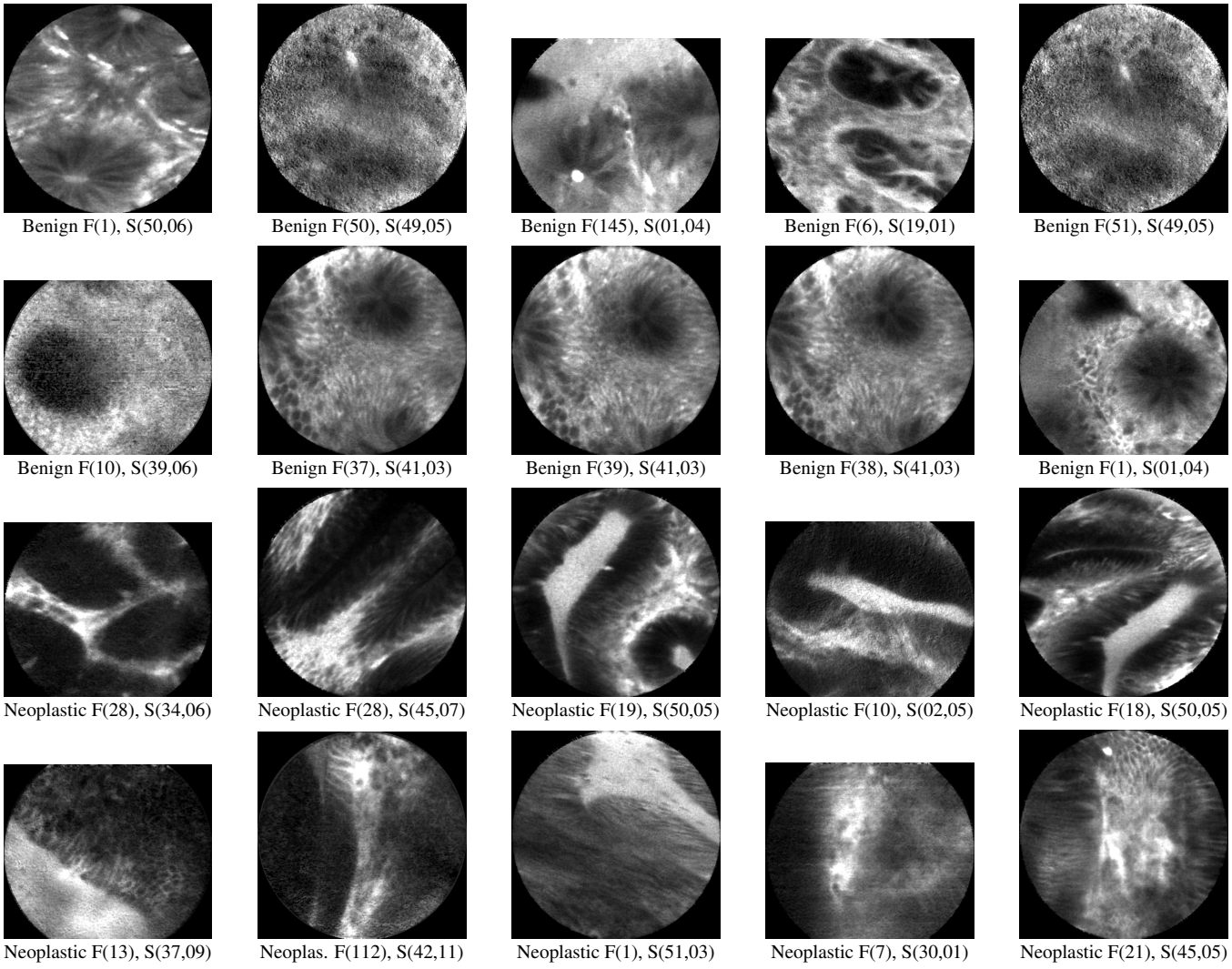

Fig. 2: Four rows of similar pCLE images provided by our method. From left to right on each row: the tested image, and its first, second, third and fourth most similar images. F indicates the frame number of the image and S the number of the video sequence. Field of view of the images: 240 μm.

sonable overlap between adjacent regions. The number K of visual words provided by the K-Means clustering was selected among values from 30 to 1500 given by the literature: the value $K = 100$ yielded satisfying classification results. The K' most discriminant visual words are selected by applying on their discriminative power a hand-picked threshold, $\theta = 0.25$, which gives good classification results when testing the whole training set without cross-validation. This threshold θ is applied inside the cross-validation loop to select a certain number of discriminant visual words, and the mean value of K' for all cross-validations is 40. Finally, the value chosen for the number k of nearest neighbors is the global value maximizing the classification accuracy on the graph of accuracy rates presented in Fig. 3 (where the best result is reached for $k = 42$). Some illustrative examples of the resulting image similarities are shown in Fig. 2.

To compare our approach with other methods, global as well as local, the validation scheme consists in the following non-biased classification: the k nearest images in the training set are retrieved, with training images not belonging to the video sequence of the image being tested (i.e. leave-n-out cross-validation, where n is the number of frames in the video of the tested image); then the votes of the closest image signatures are weighted by their distance to the tested signature. For performance comparison, the following methods are taken as references: the sparse scale invariant SIFT method [4], the statistical approach of Haralick features [6, 7] and the texture retrieval method of Textons [8]. The Haralick method computes global statistics from the co-occurrence matrix, such as contrast, correlation or variance, so as to represent an image by a vector of statistical features; this method is worth being compared with, because of its global scope. The last reference approach, the Textons method, defines for each image pixel p a "texton", as the response of a patch centered on p to a texture filter which is composed of orientation and spatial-frequency selective linear filters. While only tex-

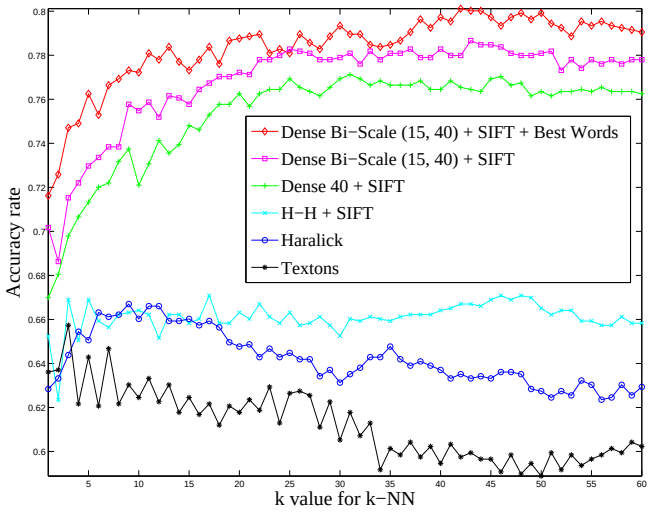

Fig. 3: Accuracy rates with leave-n-out cross-validation.

Method / Rates	Acc.	Sens.	Spec.
Dense Bi-Scale (15, 40) + SIFT + Best Words	80.1	79.5	80.8
Dense Bi-Scale (15, 40) + SIFT	78.0	75.6	80.6
Dense 40 + SIFT	76.5	71.1	82.7
Dense 15 + SIFT	70.7	75.1	65.7
H-H + SIFT	66.7	73.4	59.2
Haralick	63.5	67.0	59.6
Textons	59.8	47.3	73.7

Fig. 4: Classification results (accuracy, sensitivity, specificity) with leave-n-out cross-validation for $k = 42$ nearest neighbors.

ture information is extracted by this method, the fact that its extraction procedure is dense makes it interesting for method comparison.

After the classification process has been applied, accuracy, sensitivity and specificity rates are computed for each method, as shown in Fig. 4. According to these values, our method is the most efficient on the training database, with an accuracy rate of 80.12%, which is 13.42 points better than the standard SIFT method: the gain of accuracy can be decomposed in 9.85 points for the choice of a dense detector, 1.45 points for the bi-scale SIFT description and 2.12 points for the supervised selection of the "best" visual words. These results point out that, for our training data, the standard SIFT method, the Textons method and the Haralick method are clearly less efficient than our approach.

5. CONCLUSION

To our knowledge this is the first attempt to classify endomicroscopic images by adapting a recent and powerful local image retrieval method, the Bag of Visual Words method, introducted for recognition problem in computer vision. Although our study has been focused on one relatively small training set of pCLE images of colonic polyps, the classification results show that our objectives have been successfully reached by the methodology, which consists first in densely collecting feature information, secondly in describing this information at two different scales, one corresponding to microscopic features (cells) and the other to mesoscopic features (groups of cells), and in selecting the most discriminative visual words. Besides, our generic framework could be reasonably extended to many other image retrieval applications.

As for future work, efforts will be made to strengthen the validation of our method, for example by using ROC curves to better estimate the parameters. Concerning the classification procedure, probabilistic models with hidden variables [9] could be more efficient than the voting algorithm. Moreover, the classification results would be improved if a larger training database was considered, where all the characteristics of the image classes are better represented. Improvements could also be made for the treatment of outliers in similar images by including the temporal dimension of video sequences, which would address at the same time the problem of noise, motion distortions or partially visible macroscopic features (e.g. crypts) in a given frame. A more complete study focused on content based image retrieval could take into account, in the image description, spatial relationships between the salient regions to close the gap between highly local image features and more global image features.

6. REFERENCES

[1] 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, June 2007.

[2] J. Sivic and A. Zisserman, "Efficient visual search for objects in videos," *Proc. IEEE*, vol. 96, pp. 548–566, Apr. 2008.

[3] A.M. Buchner, M.S. Ghabril, M. Krishna, H.C. Wolfsen, and M.B. Wallace, "High-resolution confocal endomicroscopy probe system for in vivo diagnosis of colorectal neoplasia," *Gastroenterology*, vol. 135, no. 1, pp. 295, July 2008.

[4] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, Nov. 2004.

[5] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.J. Van Gool, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, pp. 43–72, Nov. 2005.

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

[7] S. Srivastava, J.J. Rodriguez, A.R. Rouse, M.A. Brewer, and A.F. Gmitro, "Computer-aided identification of ovarian cancer in confocal microendoscope images," *J. Biomed. Opt.*, vol. 13, no. 2, pp. 024021, March/April 2008.

[8] 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, June 2001.

[9] E. Hörster, T. Greif, R. Lienhart, and M. Slaney, "Comparing local feature descriptors in pLSA-based image models," in *DAGM-Symposium*. IEEE, 2008, vol. 5096, pp. 446–455.