ENDOMICROSCOPIC VIDEO RETRIEVAL USING MOSAICING AND VISUAL WORDS

B. André[1,2], T. Vercauteren[1], A. M. Buchner[3], M. B. Wallace[4], N. Ayache[2]

(1) Mauna Kea Technologies (MKT), Paris (2) Asclepios Project-Team, INRIA Sophia-Antipolis
(3) Hospital of the University of Pennsylvania, Philadelphia (4) Mayo Clinic, Jacksonville, Florida

ABSTRACT

In vivo pathology from endomicroscopy videos can be a challenge for many physicians. To ease this task, we propose a content-based video retrieval method providing, given a query video, relevant similar videos from an expert-annotated database. Our main contribution consists in revisiting the Bag of Visual Words method by weighting the contributions of the dense local regions according to the registration results of mosaicing. We perform a leave-one-patient-out k-nearest neighbors classification and show a significantly better accuracy (e.g. around 94% for 9 neighbors) when compared to using the video images independently. Less neighbors are needed to classify the queries and our signature summation technique reduces retrieval runtime.

Index Terms— Endomicroscopy, Bag of Visual Words (BVW), Leave-One-Patient-Out (LOPO), Mosaicing

1. INTRODUCTION AND CLINICAL CONTEXT

Probe-based confocal laser endomicroscopy (pCLE) enables dynamic imaging of tissue *in vivo*, at microscopic level with a miniprobe, and in real time during ongoing endoscopy. The still evolving semantic of the acquired pCLE videos remains complex for the physicians, for instance to differentiate benign and neoplastic (i.e. pathological) tissues of colonic polyps. Indeed, in these image sequences, there is a large variability of the appearance of polyp tissues having a given pathology. Besides, a video of a neoplastic polyp in the colon may also contain benign tissue. Expert physicians pointed out that the field-of-view (FOV) of single still images may not be large enough to make a robust diagnosis. The temporal dimension of videos is thus needed by the endoscopist to relate each image to its spatial context for the interpretation of partially visible macroscopic features. To aid the endoscopist in establishing a diagnosis, we propose to provide several videos that have a similar appearance to the video of interest but that have been previously annotated by expert physicians with a textual diagnosis, and potentially validated against ground truth information such as a real biopsy. In [1], we presented a content-based image retrieval (CBIR) method and successfully applied it for a pCLE database of colonic polyps. For the images of this database, we obtained better classification results than those reached by the state-of-the-art methods reviewed by [2]. In this study, the same database has been enriched. We now aim at retrieving videos and not isolated images, by revisiting the Bag of Visual Words (BVW) method explained in Section 2 and by analyzing the impact of including spatial overlap between time-related images. Our content-based video retrieval (CBVR) contributions are presented in Section 2. We first exploit the registration results of the mosaicing technique of [3] to weight the contribution of each image region to its visual word, according to the overlap of this region with the other regions densely distributed in the images. Then, we compute the video signature with a signature summation technique. This reduces both retrieval runtime and training memory. To quantify the relevance of video retrieval, we perform a k-nearest neighbors (k-NN) classification with leave-one-patient-out (LOPO) cross-validation and compare our method with other methods. In Section 4, the classification results show that our approach achieves a better accuracy on the video database. Using video data improves the results in a statistically significant manner when compared to using the images independently. Moreover, fewer nearest neighbors are necessary to classify the query at a given accuracy, which is clinically relevant for the endoscopist, who will typically examine a reasonably small number of videos.

2. VISUAL WORDS FOR IMAGE RETRIEVAL

Let us first focus on the BVW method presented in [1] to retrieve images. As the probe is translating and rotating along the tissue surface, we aim at describing pCLE images in an invariant manner with respect to translation and rotation. As the rate of fluorescein which is administrated before imaging procedure is decreasing through time, we want this description to be also invariant by any affine transformation of intensity. To this purpose, the standard BVW [2] method appeared to be the most appropriate since it extracts a local image description invariant with respect to illumination changes and some viewpoint changes, e.g., translations, rotations and scaling.

The BVW retrieval process can be decomposed into four steps: region detection, description, clustering and similarity measuring. In pCLE images, discriminative information is densely distributed over their entire field. To extract all the information, we decided to apply a dense detector made of

overlapping disks. These disk regions are localized on a dense regular grid. At the description step, the SIFT [4] descriptor computes, for each salient region, its gradient histogram vector, the gradient orientations being normalized with respect to the principal orientation of the salient region. Note that with the dense detection, the SIFT vectors are scale-dependent so the image description is not invariant by scaling. In fact, this is beneficial for our application because in colonic polyps, mesoscopic crypts and microscopic goblet cells both have a rounded shape, but are different objects characterized by their different size. To reduce the dimension of the description space, a standard K-Means clustering step builds K clusters, i.e. K visual words, from the union of the description vector sets gathered across all the images of the training database. Each description vector counts for one visual word, and an image is represented by a signature of size K which is its histogram of visual words, normalized by the number of its local regions. Given these image signatures, the distance between two images is typically defined as the χ^2 distance between their signature and the most similar training images to the image of interest are the closest ones.

3. CONTRIBUTIONS FOR VIDEO RETRIEVAL

On a still pCLE image, some discriminative patterns, e.g. an elongated crypt, may be too partially visible to characterize the pathology. Indeed, in CBIR results, we often observe still images with a very similar appearance but attached to contradictory diagnoses. For this reason, we investigated CBVR methods to retrieve similar videos instead of single images.

To address this problem, the temporal dimension of pCLE videos needs to be exploited, by including in the retrieval process the possible spatial overlap between the images from the same video sequence. In our pCLE video database, the dynamic motion within the tissue can be neglected when compared to the global motion of the probe sliding along the tissue surface. Successive images from the same video are only related by viewpoint changes, our approach can thus take advantage of the mosaicing technique of [3] to project the temporal dimension of a video sequence onto one mosaic image with a larger FOV and of higher resolution. For an efficient video retrieval, our objective is to build a single signature for each video. Indeed, having one short signature per video not only enables a reasonable memory space to store training data, but also considerably reduces the retrieval run-time.

In a first attempt, to take advantage of the super-resolution which may be present in the fused mosaic image after non-rigid registration, we considered the mosaic images as objects of interest for the retrieval. All videos of the database were splitted into stable sub-sequences identified by expert physicians. We built mosaics on these sub-sequences and applied the dense BVW method directly on the mosaic images, see the colored visual words in the middle of Fig. 1 for an illustration. Although this method provided us quite satisfying retrieval results, it takes a long runtime for the non-rigid registration of the mosaicing process. For this reason, our second attempt investigated a more efficient method which could only use the coarse registration results of mosaicing, i.e. the translation results, computed in real time.

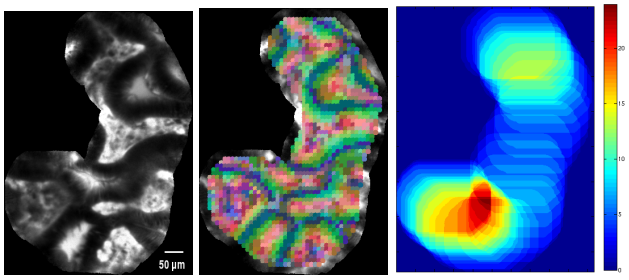

Fig. 1: From left to right: Neoplastic pCLE mosaic obtained with non-rigid registration; Colored visual words mapped to the disk regions of the mosaic image; Overlap rates of the local regions in the mosaic space, according the translation results of mosaicing.

To this purpose, we first compute independently the signatures of all the images belonging to the database of video subsequences. Then, for each sub-sequence, we use the translation results of mosaicing to build a map of the overlap rates of all local regions belonging to the images of the sequence, see Fig. 1 on the right for an illustration. To define the signature H of a video sub-sequence S, we propose to take, for each image I of the sequence, the overlap rate ρ of each region r in the image and to weight the contribution of r to the frequency of its visual word $w(r)$ by $\frac{1}{\rho}$. The visual word histogram of the video sub-sequence is given by : $H_S(w) = \frac{1}{Z}\sum_{I\in S}\sum_{r\in I}\frac{\delta(w(r),w)}{\rho(r)}$, where δ is the Kronecker notation and Z is the normalization factor. Thus, with our method called "Bag of Overlap-Weighted Visual Words" (BOWVW), we are now able to describe and retrieve stable sub-sequences of pCLE videos with a method similar to the one described in Section 2. We also define a signature for a video by considering the normalized sum of the signatures of its constitutive sub-sequences. Thanks to the BOWVW method and this histogram summation technique, the size of a video signature remains equal to the number of visual words, which reduces both retrieval runtime and training memory.

4. RESULTS AND DISCUSSION

At the Mayo Clinic in Jacksonville, the Cellvizio® system was used to image colonic polyps during surveillance colonoscopies in 68 patients. For each patient were performed one or more acquisitions of pCLE videos, one video corresponding to one particular polyp of the patient. In each of these videos, stable sub-sequences were identified by clinical experts to attach a diagnosis to the video : they differentiate pathological patterns 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 resulting database is composed of 121 videos (36 benign, 85 neoplas-

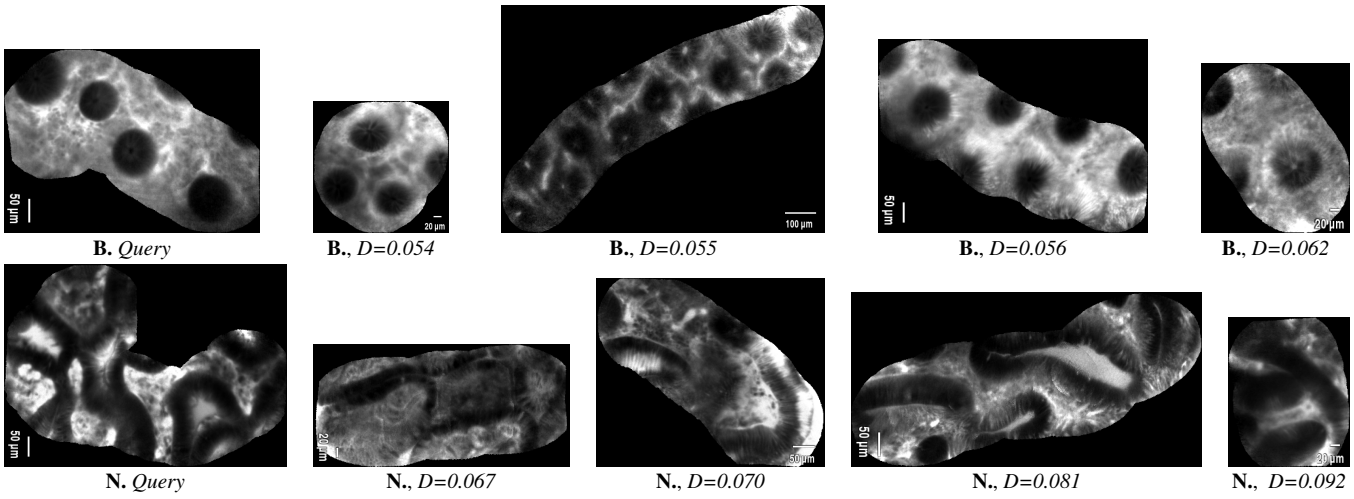

Fig. 2: Two rows of similar pCLE video sub-sequences, represented as mosaics and retrieved with the LOPO Weighted-ImOfMos method. **B.** indicates Benign and **N.** Neoplastic. From left to right on each row: the queried video sub-sequence, and its first, second, third and fourth most similar video sub-sequences, along with their similarity distance D.

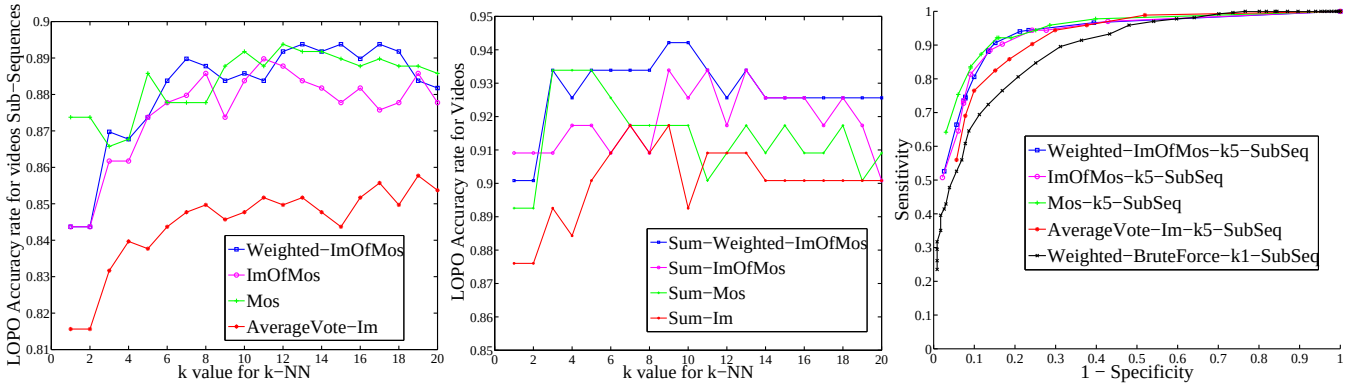

Fig. 3: From left to right : LOPO classification of pCLE video sub-sequences with $\theta = 0$; LOPO classification of pCLE videos with $\theta = 0$; ROC curves with $\theta \in [-1, 1]$ and $\theta_{BF} \in [1.0, 1.1]$, for the LOPO classification of pCLE video sub-sequences by the methods.

tic) and 499 video sub-sequences (231 benign, 268 neoplastic), leading to 4449 endomicroscopic images with a FOV of 240 μm (2292 benign, 2157 neoplastic). For all the videos of the database, the pCLE diagnosis, either benign or neoplastic, is the same as the "gold standard" established by a pathologist after histological review.

On this database, we tested several methods to retrieve image sequences. All these methods share the same BVW methodology for the image description part, with the following parameters : we considered disk regions of radius 60 pixels to cover groups of cells; we then chose 20 pixels of grid spacing to get a reasonable overlap between adjacent regions and to be nearly invariant by translation; among the values found in the literature, we took $K = 100$ visual words, which yields satisfying retrieval results. Note that the number of visual words could be reduced or their frequency could be weighted by their discriminative power w.r.t. the two classes, benign and neoplastic. However, in this study we merely aim at analyzing the impact, on the retrieval performance, of weighting the visual word frequencies. Retrieval results of our BOWVW method applied on pCLE sub-sequences can be qualitatively appreciated in Fig. 2 for a benign query and a neoplastic query. In fact, it is very difficult to have a ground-truth for content-based data retrieval. This is the reason why we chose classification as a means to quantify the relevance of the similarity results, by considering two classes, the benign class and the neoplastic class. This will also allow us to compare the performance of several retrieval methods. More precisely, we used a nearest neighbors classification 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. Thus the closest images are the most determinant. To ensure an unbiased classification, we performed a leave-one-patient-out (LOPO) cross-validation : all videos from a given patient are excluded from the training set before the training in order to be tested as queries of our retrieval and classification methods.

For the classification of video sub-sequences, we call : "Weighted-ImOfMos" the method using the BOWVW technique; "ImOfMos" the same method without overlap weight-

ing ($\rho = 1$); "Mos" the method describing the single fused mosaic image obtained with non-rigid registration; "AverageVote-Im" the method describing all the images independently and averaging their individual votes. For the classification of the whole videos, the prefix "Sum-" means that we extended the methods with the signature summation technique to retrieve videos as entities; "Sum-Im" is the method summing all the individual image signatures of the video.

In the classification accuracies shown in Fig. 3 we observe that, for the classification of videos and from $k = 3$ neighbors, the "Sum-Weighted-ImOfMos" method has an accuracy which is better than the one of "Sum-Im", and equal or better than the one of 'Sum-ImOfMos" or "Sum-Mos" methods. Moreover, when comparing the methods for the classification of video sub-sequences, the number of classified data is sufficient to perform a McNemar's test which shows that, from $k = 3$, the accuracy of the "Weighted-ImOfMos" method is statistically better than the one of "AverageVote-Im", e.g. for $k = 3$ neighbors the p-value is equal to 0.021. The best video classification results that is observed before 10 neighbors is achieved by "Sum-Weighted-ImOfMos" for $k = 9$, with an accuracy of 94.2%, a sensitivity of 97.7% and a specificity of 86.1%. For a smaller number of neighbors, "Sum-Weighted-ImOfMos" already achieves a quite satisfying accuracy, e.g. 93.4% for 3 neighbors.

When considering k nearest neighbors for a query, we computed the value of the weighted sum of their votes (-1 for benign class, $+1$ for neoplastic class) according to their proximity to the query, and we compared this value with an absolute threshold θ to classify the query as benign or neoplastic. The accuracy results were obtained with $\theta = 0$. In Fig. 3, we also draw ROC curves corresponding to the video subsequence classification results reached by the tested methods at a fixed number of neighbors $k = 5$, but with several values of the threshold $\theta \in [-1, 1]$. The closer θ is to -1 (resp. $+1$), the more weight we give to the neoplastic votes (resp. the benign votes) and the larger the sensitivity (resp. the specificity) is. We notice that, for each method, there exists a clear accuracy peak at a negative value of θ, e.g. at $\theta = -0.6$ for the "Weighted-ImOfMos" method. This illustrates the fact that global neoplastic features are more discriminative than the benign ones.

One may reproach our methodology for using an arbitrary number of visual words used for clustering and thus being dependent on the clustering results. This is the reason why we decided to compare it with a simple and efficient image classification method presented in [5], which uses no clustering. This method, referred as "BruteForce", consists in computing in the description space, for each local region of the query, its distances respectively to the closest region of the benign and pathological training data sets. If the sum of the benign distances D_B is smaller than the sum of the neoplastic distances D_N, the query is classified as benign, otherwise as neoplastic. We can easily extend the "BruteForce" method to a "Weighted-BruteForce" method for video classification, by weighting the closest distance computed for each region by the inverse of its overlap rate. Besides, a ROC curve can be obtained by introducing a multiplicative threshold $\theta_{BF} \geq 1$, and by classifying the query as neoplastic if and only if $D_N < \theta_{BF} D_B$. The ROC curve of "Weighted-BruteForce" method shows statistically worse results in comparison to the other methods, with p-values less than 0.05 with the McNemar's test. Besides, the best classification accuracies of "Weighted-BruteForce" are reached for $\theta_{BF} = 1.017 > 1$, which is confirming that local neoplastic features are also more discriminative than the benign ones. In fact, putting more weight on neoplastic patterns leads to increase the classification sensitivity, which is clinically important since it reduces the rate of false negatives.

5. CONCLUSION

Using the registration results of a mosaicing technique on endomicroscopic video data allowed us to provide the physicians with relevant annotated videos, similar to the video of interest to potentially support diagnostic decision. When compared to learning and retrieving images independently, our video retrieval method called "Bag of Overlap-Weighted Visual Words" improves the results of video classification in a statistically significant manner. Besides, it is based on histogram summations that considerably reduce both retrieval runtime and training memory. For future work, improvements could be made by taking into account the spatial relationship between local features in order to extract the spatial organization of cells. A more complete approach would describe the local $2D + t$ volumes contained in the videos to work on more accurate visual words and better combine spatial and temporal information.

6. REFERENCES

[1] B. André, T. Vercauteren, A. Perchant, M. B. Wallace, A. M. Buchner, and N. Ayache, "Endomicroscopic image retrieval and classification using invariant visual features," in *Proc. ISBI'09*, 2009, pp. 346–349.

[2] 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.

[3] 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.

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

[5] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. CVPR'08*, 2008, pp. 1–8.