# An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis

Barbara André[1,2], Tom Vercauteren[1], Anna M. Buchner[3], Muhammad Waseem Shahid[4], Michael B. Wallace[4], and Nicholas Ayache[2]

[1]Mauna Kea Technologies, Paris [2]INRIA - Asclepios, Sophia-Antipolis [3]Hospital of the University of Pennsylvania, Philadelphia [4]Mayo Clinic, Jacksonville, Florida

**Abstract.** Learning medical image interpretation is an evolutive process that requires modular training systems, from non-expert to expert users. Our study aims at developing such a system for endomicroscopy diagnosis. It uses a difficulty predictor to try and shorten the physician learning curve. As the understanding of video diagnosis is driven by visual similarities, we propose a content-based video retrieval approach to estimate the level of interpretation difficulty. The performance of our retrieval method is compared with several state of the art methods, and its genericity is demonstrated with two different clinical databases, on the Barrett's Esophagus and on colonic polyps. From our retrieval results, we learn a difficulty predictor against a ground truth given by the percentage of false diagnoses among several physicians. Our experiments show that, although our datasets are not large enough to test for statistical significance, there is a noticeable relationship between our retrieval-based difficulty estimation and the difficulty experienced by the physicians.

## 1 Introduction

**Objective** The understanding of pathologies through the analysis of image sequences is a subjective learning experience which may be supported by modular training systems. Particularly, the early diagnosis of epithelial cancers from *in vivo* endomicroscopy is a challenging task for many non-expert endoscopists. Our objective is to develop a modular training system for endomicroscopy diagnosis, by adapting the difficulty level according to the expertise of the physician.

The training simulator, illustrated in Fig. 1 on the top right, consists in a quiz. Given a level of difficulty, a pool of endomicroscopic videos whose average difficulty matches the current level is randomly chosen from the set of the training videos. By iterating this process with increasing levels of interpretation difficulty, the physician may be able to learn faster.

**State of the art in estimating interpretation difficulty** Typical studies on query difficulty estimation consider textual queries, and not image queries. Besides, they usually do not predict the difficulty of the query interpretation

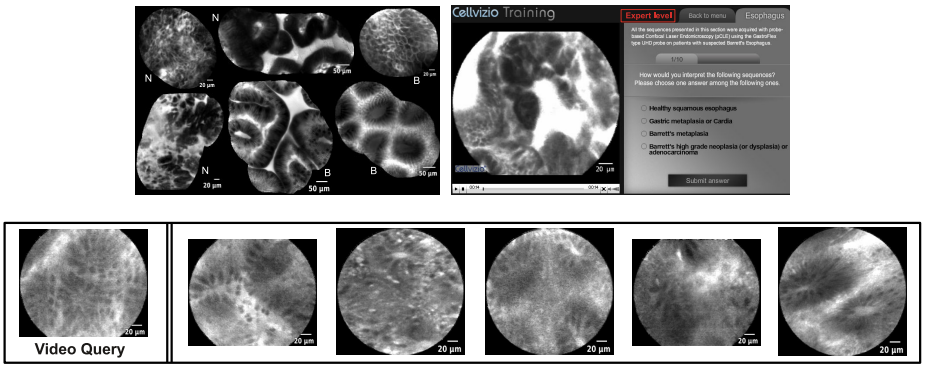


**Fig. 1:** Top left: 6 mosaic images of *Barrett* (**B**: Benign, **N**: Neoplastic). Top right: Screenshot of `www.cellvizio.net` Self-Learning Tool, with the added difficulty level information. Bottom: Retrieval example of our CBVR method on *Colon* with the blinded query video (left) and its 5 most similar videos represented by single frames.

but rather the *performance* of the query in order to estimate the quality of its retrieval results. However, given the tight analogy between text retrieval and image retrieval, the difficulty criteria used by these methods, most of which were presented in a survey by Hauff et al. [1], may also be useful for our study. In particular, Zhao et al. [2] estimated the performance of a textual query from similarity scores, but also from term frequency - inverse document frequency (TF-IDF) weights [3] extracted during the indexing time. In all these studies, the predictor validation process takes as ground truth an indicator of the performance of the retrieval system, such as the Average Precision (AP). Nevertheless, Scholer and Garcia [4] demonstrated that the correlation between the estimated difficulty and the measured retrieval performance highly depends on the chosen retrieval system. Considering human performance in rating x-ray images as a ground truth, Schwaninger et al. [5] proposed a statistical approach to estimate the image query difficulty solely from image measurements. Turpin and Scholer [6] highlighted the fact that it is not easy to establish, for simple tasks like instance recall or question answering, a significant relationship between human performance and the performance of a retrieval system that uses precision-based measures to predict the query difficulty.

For our study, we consider videos as queries. We propose to learn a query difficulty predictor using relevant attributes from a Content-Based Video Retrieval (CBVR) method. We have two types of ground truths. For video retrieval, a diagnosis ground truth is the set of histological diagnoses of the biopsies associated to all the videos of the database. For interpretation difficulty, a difficulty ground truth is given by the percentage of false video-based diagnoses among several physicians on a subset of the video database. Histological diagnosis and video-based diagnosis both consist in differentiating benign from neoplastic (i.e.

pathological) lesions. In these conditions, we aim at establishing a relationship between the physicians performance and our predictor.

**Materials** Probe-based confocal laser endomicroscopy (pCLE) allows the endoscopist to image the epithelial surface *in vivo*, at microscopic level with a miniprobe, and in real-time (12 frames per second) during an ongoing endoscopy.

The first pCLE database is of colonic polyps videos acquired by physicians at the Mayo Clinic in Jacksonville, Florida, USA. 68 patients underwent a surveillance colonoscopy with pCLE for fluorescein-aided imaging of suspicious colonic polyps before their removal. For each patient, pCLE was performed of each detected polyp with one video corresponding to each particular polyp. Our resulting *Colon* database is composed of 121 videos (36 benign, 85 neoplastic) split into 499 stable video sub-sequences (231 benign, 268 neoplastic). 11 endoscopists, among whose 3 experts and 8 non-experts, individually established a pCLE diagnosis on 63 videos (18 benign, 45 neoplastic) of the database. On the non-expert diagnosis database, interobserver agreement was assessed in the study of Buchner et al. [7], with an average accuracy of 72% (sensitivity 82%, specificity 53%). On the expert diagnosis database, Gomez et al. [8] showed an interobserver agreement with an average accuracy of 75% (sensitivity 76%, specificity 72%). Thus, although pCLE is relatively new to many physicians, the learning curve pattern of pCLE in predicting neoplastic lesions was demonstrated with improved accuracies in time as observers' experience increased.

The second pCLE database is related to a different clinical application, namely the Barrett's Esophagus, and was provided by the multicentric "DONT BIOPCE" [9] study (Detection Of Neoplastic Tissue in Barrett's esophagus with In vivO Probe-based Confocal Endomicroscopy). Our resulting *Barrett* database includes 76 patients and contains 123 videos (62 benign, 61 neoplastic) split into 862 stable video sub-sequences (417 benign, 445 neoplastic). 21 endoscopists, among whose 9 experts and 12 non-experts, individually established a pCLE diagnosis on 20 videos (9 benign, 11 neoplastic) of the database.

For all these training videos, the pCLE diagnosis, either benign or neoplastic, is the same as the *gold standard* established by a pathologist after the histological review of biopsies acquired on the imaging spots.

## 2 Estimating the interpretation difficulty

For difficulty estimation, our ground truth is given by the percentage, for each query video, of false diagnoses among the physicians.As the understanding of video diagnosis by the physicians is driven by the observation of visual similarities between the query video and training videos, it makes sense to predict the query difficulty based on similarity results of video retrieval.

To learn a difficulty predictor, our idea is to exploit, as relevant attributes, the results of our video retrieval method applied to the training database. Potential relevant attributes are the class $c_q \in \{-1, +1\}$ of the video query $q$, the classes $c^{i \in \{1,k\}} \in \{-1, +1\}$ of its $k$ nearest neighbors and the similarity distances $\delta^{i \in \{1,k\}}$

to them. Given the small number of videos tested by the involved physicians, too many attributes for difficulty learning may lead to over-fitting. For this reason, we decided to extract one efficient and intuitive difficulty attribute $\alpha$ from the retrieval results. For each query video, we considered the retrieval error between the average of the neighbors' votes and the class of the query: $\alpha_q = 1 - c_q(\sum_{i\in\{1,k\}} c^i w_{c^i})/(\sum_{i\in\{1,k\}} w_{c^i})$ , where $w_{-1} = 1$ and $w_{+1}$ is a constant weight applied to the neoplastic votes. Introducing $w_{+1}$ allows us to take into account the possible emphasis of neoplastic votes with respect to the benign votes. Our query difficulty predictor $P$ is thus defined as $P(q) = \alpha_q$ for each query video $q$. Its relevance can be evaluated by a simple correlation measure between the estimated difficulties of all tested videos and their ground truth values. In this case, as there is no learning process, cross-validation is not necessary.

## 3 Video retrieval method

As one of the most popular method for Content-Based Image Retrieval (CBIR), the Bag-of-Visual-Words (BoW) method presented by Zhang et al. [10] aims at extracting a local image description that is both efficient to use and invariant with respect to viewpoint changes and illumination changes. Its methodology consists in first finding and describing salient local features, then in quantizing them into $K$ clusters named visual words, and in representing the image by its signature which is the histogram of visual words.

As the field-of-view (FOV) of single images may sometimes be insufficient to establish a diagnosis, we revisited in [11] the standard BoW method to retrieve videos, and not only single images. We consider each video as a set of stable sub-sequences corresponding to a relatively smooth movement of the pCLE probe along the tissue surface. We then use a video-mosaicing technique to project the temporal dimension of each sub-sequence onto one mosaic image with a larger FOV and of higher resolution. Thus, each video is a set of mosaic images.

We adapt the BoW method to retrieve endomicroscopic mosaic images, some of which are shown in Fig. 1 on the top left. In colonic polyps, a mesoscopic crypt and a microscopic goblet cell both have a rounded shape, but are different objects characterized by their different sizes. Our description should thus not be invariant with respect to scaling. Noticing that discriminative information is densely distributed in pCLE mosaic images, we decided to apply a dense detector made of overlapping disks of constant radius on a regular grid. For the description step, we used the Scale Invariant Feature Transform (SIFT) descriptor, whose combination with our dense detector keeps all the BoW related invariants except scale invariance. After mosaic image description, we define the signature of a video as the normalized sum of the visual word histograms associated to each of the constitutive mosaic images. Fig. 1 on the bottom shows a CBVR example, where the retrieved videos are not only similar but also unblinded, i.e. displayed along with their contextual and diagnostic information.

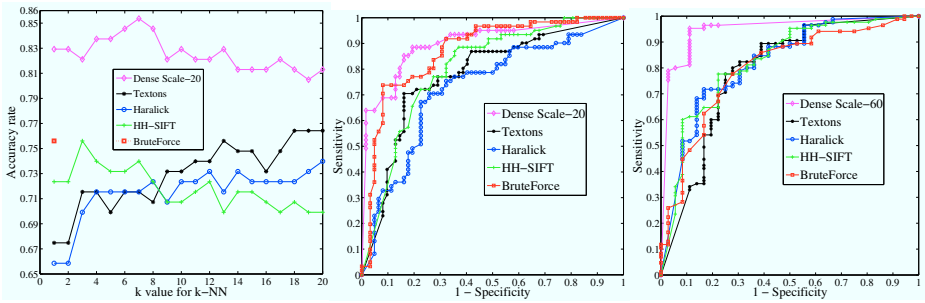


**Fig. 2:** Left: Method comparison for the LOPO classification of pCLE videos with $\theta = 0$ on *Barrett*. Middle and right: Corresponding ROC curves at $k = 5$ neighbors, on *Barrett* (middle) and on *Colon* (right).

## 4 Evaluation of the relevance of the retrieval method

Evaluating the relevance of retrieval results is a difficult problem. Because of the subjective appreciation of visual similarities, it is quite difficult to have a ground-truth. A simple evaluation method is to perform a classification based on the retrieval results and to estimate its accuracy. As our focus is on retrieval and not on classification, we chose one of the most straightforward classification method, the $k$-NN method, even though any other method could be easily plugged in. We consider two classes, benign (vote $= -1$) and neoplastic (vote $= +1$). When considering $k$ neighbors for a query, we compute the value of the weighted sum of their votes according to their similarity distance to the query, and we compare this value with an absolute threshold $\theta$ to classify the query as benign or neoplastic. $\theta$ can be used to arbitrate between sensitivity and specificity.

In the current work, we apply for the first time our video retrieval method to two different pCLE databases, *Colon* and *Barrett*. Given their relatively small sizes, we use for each of them the whole database both for training and testing. If we only perform a leave-one-out cross-validation, the independence assumption is not respected because there are several videos acquired from the same patient. Since this may cause bias, we chose to perform a leave-one-patient-out (LOPO) cross-validation: all videos from a given patient are excluded from the training set in order to be then tested as queries of our retrieval and classification methods.

For method comparison, we will take as references the following CBIR methods, which we extended to CBVR by applying our signature summation technique: the HH-SIFT method presented by Zhang et al. [10] a sparse detector, the standard approach of Haralick features, the texture retrieval Textons method of Leung and Malik [12], and an efficient image classification method presented by Boiman et al. [13], referred as "BruteForce", that uses no clustering.

For our retrieval method, we considered disk regions of radius 60 pixels for the *Colon* database, and 20 pixels for the *Barrett* database whose discriminative patterns appear at a finer scale. We then chose 20 pixels of grid spacing to get a

reasonable overlap between adjacent regions and thus be nearly invariant with respect to translation. For the number $K$ of visual words provided by the $K$-Means clustering, among the values from 10 to 30000 in the literature, we chose the value $K = 100$ whose performance appeared to be sufficient for our needs.

The accuracy results of video classification on *Barrett* are presented in Fig. 2 on the left. In agreement with the presented ROC curves, the accuracy results obtained on *Colon* are even better. Our retrieval method outperforms all the compared methods with a gain of accuracy greater than 12 percentage points (pp.) on *Colon*, and greater than 9 pp. on *Barrett*. McNemar's tests show that, when the number $k$ of neighbors is fixed, the improvement of our method with respect to all others is statistically significant: $p$-value $< 0.011$ for $k \in [1, 10]$ on *Colon* and $p$-value $< 0.043$ for $k \in [1, 2] \cup [4, 8]$ on *Barrett*. This shows the genericity of our retrieval method, which is successfully applied to two different clinical application, with: 93.4% of accuracy (sensitivity 95.3%, specificity 88.9%) at $k = 3$ neighbors on the *Colon* database, and 85.4% of accuracy (sensitivity 90.2%, specificity 80.7%) at $k = 7$ neighbors on the *Barrett* database.

## 5 Results of the difficulty estimation method

**Results on the *Barrett* database.** We experimented our difficulty predictor presented in Section 2 on *Barrett*. The best correlation results were obtained with $k = 10$ neighbors and a neoplastic weight $w_{+1} = 0.4$. The correlation values reach 0.78 when learning from the subset of videos diagnosed by all the physicians, 0.63 (resp. 0.80) when learning only from the experts (resp. the non-experts). The corresponding joint histogram is presented in Fig. 3 on the top, along with the histogram of the difficulty ground truth values. We observe a noticeable relationship between ground truth and our proposed difficulty estimation, which confirms the efficiency of our retrieval-based attribute for intuitive difficulty estimation.

**Perspectives for the *Colon* database.** On *Colon* the difficulty estimation results are not as good as on *Barrett*. With $k = 10$ neighbors and a neoplastic weight $w_{+1} = 6$, the correlation values reach 0.45 when learning from the subset of videos diagnosed by all the physicians, 0.30 (resp. 0.45) when learning from experts (resp. non-experts). In order to improve these results, we propose to investigate a machine learning-based approach, which will need more relevant attributes. As the video dataset for which we have the difficulty ground truth is relatively small, we decided to add one discriminative power attribute $\beta$ and to learn the difficulty predictor from the two attributes $\alpha$ and $\beta$ by using a robust linear regression model. Our discriminative power attribute $\beta$ reflects the deviation of the "signed" discriminative power of the query signature, with respect to the benign and the neoplastic classes: $\beta = std(\sum_{i \in \{1,K\}} f_i d(i))$. In this formula, the visual word $i$ has a frequency $f_i$ in the video query and its "signed" discriminative power $d(i)$ is given by the adapted Fisher criterion: $d(i) = (\mu_{-1} - \mu_{+1})|\mu_{-1} - \mu_{+1}|/(0.5\ (\sigma^2_{-1} + \sigma^2_{+1}))$, where $\mu_c$ and $\sigma_c$ are respectively the mean and the variance of the frequency distribution of the visual word

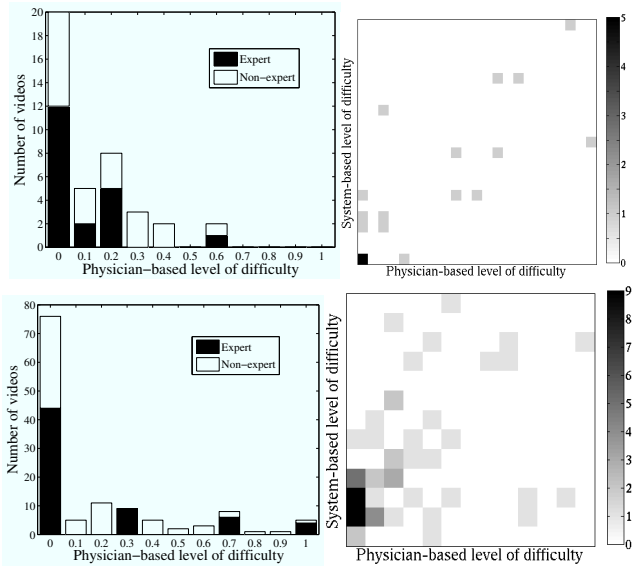


**Fig. 3:** Left: Difficulty ground truth histograms on *Barrett* (top) and on *Colon* (bottom). Right: Joint histograms on *Barrett* (top) and on *Colon* (bottom); $x$-axis is the difficulty of all the physicians and $y$-axis is our estimated difficulty. On *Barrett*, 21 physicians, 9 expert and 12 non expert, individually diagnosed 20 videos. On *Colon*, 11 physicians, 3 expert and 8 non expert, individually diagnosed 63 videos.

$i$ in the videos belonging to class $c$. The correlation values obtained by the robust linear regression model with cross-validation reach 0.48 when learning from the subset of videos diagnosed by all the physicians, 0.33 when learning only from the experts and 0.47 when learning only from the non-experts. Even if these correlation results are less convincing than those obtained on *Barrett*, the correlation tendency can be qualitatively appreciated. The corresponding joint histogram is presented in Fig. 3 on the bottom. To automate the optimal attributes selection and to explore more potentially relevant attributes for difficulty estimation, further experiments based on model selection need to be investigated, for example using the Akaike information criterion. Besides, selection criteria commonly used in active learning [14] may help to provide a better difficulty estimation.

## 6 Conclusion

To our knowledge this study proposes the first approach to estimate interpretation difficulty for endomicroscopy training, based on an original method of Content-Based Video Retrieval. Our experiments have demonstrated that there is a noticeable relationship between our retrieval-based difficulty estimation and the difficulty experienced by the physicians. Moreover, we showed the promising

genericity of our difficulty estimation method by applying it on two different clinical databases, one on the Barrett's Esophagus and the other on colonic polyps. Our method could also be potentially applied to other imaging applications.

On one hand we have the diagnosis ground truth for all the videos belonging to our two large databases, on the other hand we have the difficulty ground truth on a small subset of each database. The method proposed in this work can then be used to estimate the interpretation difficulty on the remaining videos. The complete databases could thus be used in a training simulator that features difficulty level selection. This should make endomicroscopy training more relevant. Finally, a clinical validation would be required to see whether such a structured training simulator could help shorten the physician learning curve.

## References

1. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: Proc. CIKM'08. (2008) 1419–1420
2. Zhao, Y., Scholer, F., Tsegay, Y.: Effective pre-retrieval query performance prediction using similarity and variability evidence. In: ECIR. (2008) 52–64
3. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. In: Information Processing and Management. (1988) 513–523
4. Scholer, F., Garcia, S.: A case for improved evaluation of query difficulty prediction. In: Proc. SIGIR'09. (2009) 640–641
5. Schwaninger, A., Michel, S., Bolfing, A.: A statistical approach for image difficulty estimation in X-ray screening using image measurements. In: Proc. APGV'07. (2007) 123–130
6. Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: Proc. SIGIR'06. (2006) 11–18
7. Buchner, A.M., Gomez, V., Gill, K.R., Ghabril, M., Scimeca, D., Shahid, M.W., Achem, S.R., Picco, M.F., Riegert-Johnson, D., Raimondo, M., Wolfsen, H.C., Woodward, T.A., Hasan, M.K., Wallace, M.B.: The learning curve for in vivo probe based Confocal Laser Endomicroscopy (pCLE) for prediction of colorectal neoplasia. Gastrointestinal Endoscopy **69**(5) (April 2009) AB364–AB365
8. Gomez, V., Buchner, A.M., Dekker, E., van den Broek, F.J., Meining, A., Shahid, M.W., Ghabril, M., Fockens, P., Wallace, M.B.: Interobserver agreement and accuracy among international experts of probe-based confocal laser microscopy (pCLE) in predicting colorectal neoplasia. Endoscopy **in press** (2010)
9. DONT BIOPCE: `http://clinicaltrials.gov/ct2/show/NCT00795184`.
10. 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
11. André, B., Vercauteren, T., Wallace, M.B., Buchner, A.M., Ayache, N.: Endomicroscopic video retrieval using mosaicing and visual words. In: Proc. ISBI'10. (2010) in press.
12. 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
13. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proc. CVPR'08. (2008) 1–8
14. Hoi, S.C.H., Jin, R., Zhu, J., Lyu, M.R.: Semi-supervised SVM batch mode active learning for image retrieval. In: Proc. CVPR'08. (2008) 24–26