Retrieval evaluation and distance learning from perceived similarity between endomicroscopy videos

Barbara André[1,2], Tom Vercauteren[1], Anna M. Buchner[3], 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. Evaluating content-based retrieval (CBR) is challenging because it requires an adequate ground-truth. When the available ground-truth is limited to textual metadata such as pathological classes, retrieval results can only be evaluated indirectly, for example in terms of classification performance. In this study we first present a tool to generate perceived similarity ground-truth that enables direct evaluation of endomicroscopic video retrieval. This tool uses a four-points Likert scale and collects subjective pairwise similarities perceived by multiple expert observers. We then evaluate against the generated ground-truth a previously developed dense bag-of-visual-words method for endomicroscopic video retrieval. Confirming the results of previous indirect evaluation based on classification, our direct evaluation shows that this method significantly outperforms several other state-of-the-art CBR methods. In a second step, we propose to improve the CBR method by learning an adjusted similarity metric from the perceived similarity ground-truth. By minimizing a margin-based cost function that differentiates similar and dissimilar video pairs, we learn a weight vector applied to the visual word signatures of videos. Using cross-validation, we demonstrate that the learned similarity distance is significantly better correlated with the perceived similarity than the original visual-word-based distance.

1 Introduction

Successfully developed in the field of computer vision, content-based retrieval (CBR) methods also have valuable applications in the field of medical imaging. In particular, probe-based confocal laser endomicroscopy (pCLE) is a recent imaging technology that enables the endoscopist to acquire, *in vivo*, microscopic video sequences of the epithelium. Because *in vivo* diagnostic interpretation of a pCLE video is still challenging for many endoscopists, it could be supported by the automated real-time extraction of visually similar videos that have already been annotated with a textual diagnosis. We previously developed in [1] a dense bag-of-visual-words (BoW) method for pCLE video retrieval, called "Dense-Sift", which provides qualitatively relevant retrieval results. When evaluated in terms

of classification, "Dense-Sift" was shown to significantly outperform several state-of-the-art CBR methods referred to as "competitive methods" in Section 2. However, there is a high variability in the appearance of pCLE videos, even within the same pathological class. In order to measure the adequacy of pCLE video retrieval, we propose to evaluate the "Dense-Sift" method directly in terms of visual similarity distance. To this purpose, we develop in Section 3 an online survey tool called "Visual Similarity Scoring" (VSS) to help pCLE experts in generating a perceived similarity ground-truth. In Section 4, all state-of-the-art methods are evaluated against the generated ground-truth and we show that, with statistical significance, "Dense-Sift" proves to be the best. Leveraging the perceived similarity used for evaluation purposes, we propose, in a second step, to improve the "Dense-Sift" retrieval method by learning from this ground truth an adjusted similarity metric. Our metric learning technique, presented in Section 5, is based on a visual word weighting scheme which we evaluate using cross-validation. The learned similarity metric is shown to be significantly better correlated with the perceived similarity than the original visual-word-based distance.

2 State-of-the-art in CBR and distance metric learning

Among the state-of-the-art methods in CBR, the BoW method of Zhang et al. [2], 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 [3] is based on a dense description of local texture features. Adjusting these approaches for pCLE retrieval, the "Dense-Sift" method of [1] uses a dense SIFT description. Such a description is invariant to in-plane rotations or translations changes that are due to the motion of the pCLE miniprobe in contact with the epithelium, and to the possible illumination changes that are due to the leakage of fluorescein used in pCLE. "Dense-Sift" also enables the extension from pCLE image retrieval to pCLE video retrieval by leveraging video mosaicing results. Another CBR method shown as a competitive method in [1] is the "Haralick" [4] method based on global statistical features. In this study, we choose to evaluate these four CBR methods in terms of visual similarity distance, in order to compare their retrieval performances.

Distance metric learning has been investigated by rather recent studies to improve classification or recognition methods. Yang et al. [5] 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 [6] 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. [7] 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.

Since the last approach relies on a matching ground-truth that is closer to the pairwise similarity ground-truth that we present in the next section, our proposed metric learning technique is inspired from the method of [7] and applies the transformation to the visual words signatures of videos.

3 Generation of perceived similarity ground-truth

Our video database contains 118 pCLE videos of colonic polyps that were acquired from 66 patients for the study of Buchner et al. [8]. The length of these videos is ranging from 1 second to 4 minutes. To generate a pairwise similarity ground-truth between these videos, we designed an online survey tool, called VSS [9], that allows multiple human 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 requires a lot of time, the VSS supports this task by making available the whole video content and for each video, a set of static mosaic images providing a visual summary. Indeed, Dabizzi et al. [10] recently showed that pCLE mosaics have the potential to replace pCLE videos for a comparable diagnosis accuracy and a significantly shorter interpretation time. We also paid attention to how video couples should be drawn by the VSS. If the video couples had been randomly drawn, the probability of drawing dissimilar videos would be much higher than the probability of drawing very similar videos, which would thus be poorly represented in ground-truth data. To solve this problem, we used the *a priori* similarity distance D_{prior} computed by the "Dense-Sift" method to

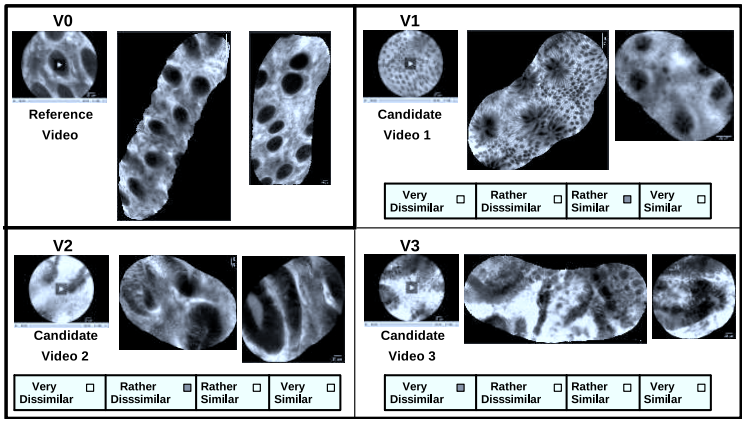

Fig. 1: Schematic outline of the online "Visual Similarity Scoring" tool showing an example of a scoring process, where 3 video couples $(V0, V1)$, $(V0, V2)$ and $(V0, V3)$ are proposed. Each video is summarized by a set of mosaic images.

draw according to the following law: the probability of drawing a video couple $(V1, V2)$ is proportional to the inverse of the density of $D_{prior}(V1, V2)$. Each scoring process, as illustrated in Fig. 1, is defined by the drawing of 3 video couples $(V0, V1)$, $(V0, V2)$ and $(V0, V3)$, where the candidate videos $V1$, $V2$ and $V3$ belong to patients that are different from the patient of the reference video $V0$, in order to exclude any patient-related bias. 17 observers, ranging from middle expert to expert in pCLE diagnosis, performed as many scoring processes as they could. The averaging time to score 3 video couples during one scoring process was 10 minutes. Our generated ground-truth can be represented as an oriented graph $G = (V, E)$ where the nodes V are the videos and where each couple of videos may be connected by zero, one or multiple 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, 3804 similarity scores were given on 1951 distinct video couples. Only 14.5% of all 13434 distinct video couples were scored. Although the similarity graph is very sparse, we demonstrate in the following sections that it constitutes a first valuable ground-truth, not only for retrieval evaluation but also to learn an adjusted similarity distance.

4 Evaluation of CBR methods against ground-truth

The evaluation of a CBR method against ground-truth can be qualitatively illustrated by the four superimposed histograms $H_L, L \in \{-2, -1, +1, +2\}$ shown in

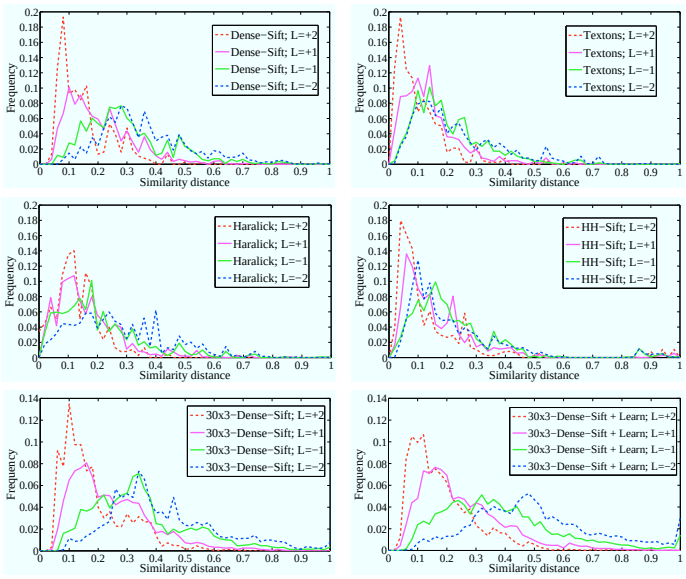

Fig. 2: Histograms of L-scored similarity distances computed by the CBR methods.

Fig. 2: H_L is the histogram of the similarity distances which were computed by the CBR method in the restricted domain of all L-scored video couples, where L is one of the four Likert points from "very dissimilar" to "very similar". We observe that these histograms are correctly ordered with respect to the four Likert points for all methods, except for "HH-Sift" that switches H_{-1} and H_{-2}. We also notice that the histograms are better separated for "Dense-Sift" than for the other methods. This is quantitatively confirmed by the histogram separability measures, given by the Bhattacharyya distance, that are shown in the supplemental material at `http://hal.inria.fr/inria-00598301/fr/`.

Possible indicators of the correlation between the visual-word-based similarity distance and the ground-truth distance are Pearson π product moment, Spearman ρ and Kendall τ. Compared to π which measures linear dependence based on the data values, ρ and τ are better adapted to the psychometric Likert scale because they measure monotone dependence based on the data ranks [11]. 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 the correlation coefficients that are associated to each CBR method, 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 perform Steiger's Z-tests, as recommended by Meng et al. [12], and we apply it to Kendall τ. The correlation results shown in Table 1 demonstrate that, with statistical significance, "Dense-Sift" is better than all other competitive methods, while "Textons" and "Haralick" are better than "HH-Sift".

Standard recall curves are a common means of evaluating retrieval performance. However, because of the sparsity of the ground-truth, it is not possible to compute them in our case. Instead, we need "sparse recall" curves. At a fixed number k, we define the sparse recall value of a CBR method as the percentage of L-scored video couples, with $L = +1$ or $+2$, for which one of the two videos has

CBIR method M	**M1** 30x3-DS-learn	**M2** 30x3-DS	**M3** Dense-Sift (DS)	**M4** Textons	**M5** Haralick	**M6** HH-Sift
Pearson π	52.8 %	46.0 %	47.8 %	32.7 %	33.8 %	15.7 %
standard error σ_{est}	0.8 %	0.9 %				
Spearman ρ	56.6 %	49.0 %	51.5 %	35.4 %	34.2 %	21.8 %
standard error σ_{est}	0.9 %	1.1 %				
Kendall τ	52.6 %	45.2 %	47.0 %	32.1 %	30.6 %	19.4 %
standard error σ_{est}	0.9 %	1.0 %				
Steiger's Z-test on τ	$>$ **M2**		$>$ **M4-M5-M6**	$>$ **M6**	$>$ **M6**	
p-value p	$p = 0.018$		$p < 10^{-4}$	$p < 10^{-4}$	$p < 10^{-4}$	

Table 1: Indicators of correlation between similarity distance computed by the CBIR methods and ground-truth. σ_{est} is the standard deviation of the estimator, it can be computed from the standard deviation of the n samples $\sigma_{samples} = \sqrt{n-1}.\sigma_{est}$. The difference between methods **M4** and **M5** is not statistically significant ($p > 0.3$).

been retrieved among the k nearest neighbors of the other video. Sparse recall curves in Fig. 3 show that "Dense-Sift" extracts similar videos much faster than the other methods in a small retrieval neighborhood, which is clinically relevant for our pCLE application. Thus, local similarity distances are better captured by the "Dense-Sift" method.

5 Distance learning from perceived similarity

As mentioned in Section 2, we now propose a metric learning technique that is inspired from the method of Philbin et al. [7]. Our objective is to transform video signatures, that are histograms of visual words, into new signatures where visual words are now weighted by a vector w that better discriminates between similar videos and dissimilar videos. We thus consider two groups: D_+ is the set of video couples that have been scored with $+2$ or $+1$, and D_- is the set of video couples that have been scored with -2 or -1. Our constraints are the following: the weights w should be positive in order to maintain the positiveness of visual word frequencies, the χ^2 metric used by standard BoW methods should

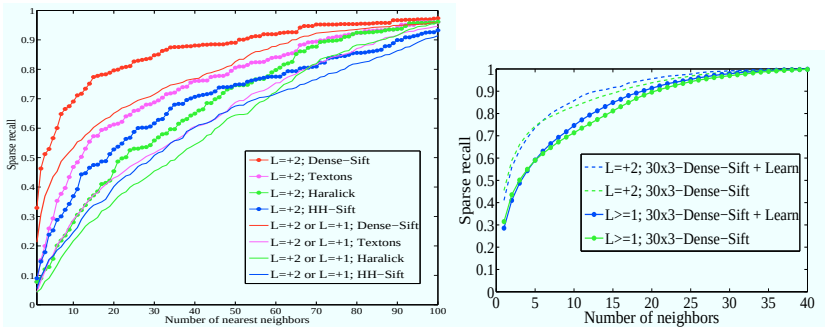

Fig. 3: Sparse recall curves associated the two methods in L-scored domains.

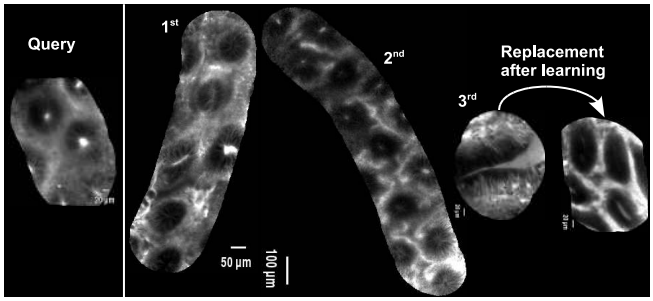

Fig. 4: Example of pCLE video query with its 3 nearest neighbors retrieved by "Dense-Sift" before and after metric learning. Videos are represented by mosaic images.

be the distance between video signatures and the new signatures should be $L1$-normalized before χ^2 distances are measured. We optimize the transformation w by minimizing the following margin-based cost function:

$$f(w) = \frac{1}{Card(D_+)} \sum_{(x,y)\in D_+} L(b - \chi^2(\frac{w \circ s_x}{||w \circ s_x||_{L1}}, \frac{w \circ s_y}{||w \circ s_y||_{L1}})) + \gamma \frac{1}{Card(D_-)} \sum_{(x,y)\in D_-} L(\chi^2(\frac{w \circ s_x}{||w \circ s_x||_{L1}}, \frac{w \circ s_y}{||w \circ s_y||_{L1}}) - b) \quad (1)$$

where b is the margin, s_x is the visual word signature of the video x, $\circ$ is the Hadamard (element-wise) product, $L(z) = log(1 + e^{-z})$ is the logistic-loss function and γ is a constant that potentially penalizes either dissimilar nearby videos or similar remote videos. The learned similarity distance is then defined by:

$$D_{learn}(x, y) = \chi^2(\frac{w_{opt} \circ s_x}{||w_{opt} \circ s_x||_{L1}}, \frac{w_{opt} \circ s_y}{||w_{opt} \circ s_y||_{L1}}) \quad (2)$$

To exclude the learning bias, we apply this distance learning technique using $m \times q$-fold cross-validation: we performed m random partitions of our database into q video subsets. Each of these subsets is successively the testing set and the union of the other subsets is the training set for both video retrieval and distance learning. To eliminate patient-related bias, all videos from the same patient are in the same subset. Given our sparse ground-truth, 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 took $m = 30$ and $q = 3$. Then, by choosing $\gamma = 5$ and $b = (\text{ median } (H^{train}_{+2,+1}) + \text{ median } (H^{train}_{-2,-1}))/2$ as an intuitive value for the margin b for each training set, we show in the following that we obtain satisfying correlation results with respect to the ground truth.

As "Dense-Sift" proved to be the best CBR method, we propose to use its visual word signatures as inputs of the learning process in order to improve its visual-word-based distance. In order to compare the performances of the learned similarity distance with those of the visual-word-based distance, "Dense-Sift" was re-trained on each training subset and re-evaluated on corresponding testing subsets. We call "30x3-fold-Dense-Sift" the cross-validated "Dense-Sift" without metric learning and "30x3-fold-Dense-Sift-learn" the same one improved by metric learning. The superimposed histograms H_L for the retrieval method before and after learning are represented in the bottom of Fig. 2. We observe that these histograms are better separated after the metric learning process, which is confirmed by the Bhattacharyya distances shown in the supplemental material.

Although the sparse recall curves of the retrieval method before and after learning are very close to each other, as shown in Fig. 3, the metric learning process globally improved the performance of the retrieval method in terms of perceived visual similarities. Indeed, the correlation results shown in Table 1 demonstrate that, with statistical significance, the learned similarity distance is better correlated than the original visual-word-based distance with the ground-truth similarity distance. Besides, for some cases as the one shown in Fig. 4, we observe that first neighbors are qualitatively more similar after metric learning.

6 Conclusion

The main contributions of this study are the generation of a valuable ground-truth for perceived visual similarities between endomicroscopic videos, the evaluation of content-based retrieval methods in terms of visual similarity and the learning of an adjusted similarity distance. The proposed methodology could be applied to other medical or non-medical databases. Our evaluation experiments confirmed that the dense BoW method for endomicroscopic video retrieval has better performances than other competitive methods, not only in terms of pathological classification but also in terms of correlation with a ground-truth similarity distance. Future work will focus on enlarging the ground truth database to investigate more sophisticated metric learning techniques. Our long-term goal is to improve endomicroscopic video retrieval and assess whether it could support the endoscopists in establishing more accurate *in vivo* diagnosis.

References

1. 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) 1419–1422
2. 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
3. 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
4. Haralick, R.M.: Statistical and structural approaches to texture. In: Proc. IEEE. Volume 67. (1979) 786–804
5. Yang, L., Jin, R., Mummert, L., Sukthankar, R., Goode, A., Zheng, B., Hoi, S.C.H., Satyanarayanan, M.: A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **32** (2010) 30–44
6. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10** (2009) 207–244
7. Philbin, J., Isard, M., Sivic, J., Zisserman, A.: Descriptor learning for efficient retrieval. In: Proc. ECCV'10. (2010) 677–691
8. Buchner, A.M., Shahid, M.W., Heckman, M.G., Krishna, M., Ghabril, M., Hasan, M., Crook, J.E., Gomez, V., Raimondo, M., Woodward, T., Wolfsen, H., Wallace, M.B.: Comparison of probe based confocal laser endomicroscopy with virtual chromoendoscopy for classification of colon polyps. Gastroenterology **138**(3) (2009) 834–842
9. Visual Similarity Scoring (VSS): `http://smartatlas.maunakeatech.com`, login: MICCAI-User, password: MICCAI2011.
10. Dabizzi, E., Shahid, M.W., Qumseya, B., Othman, M., Wallace, M.B.: Comparison between video and mosaics viewing mode of confocal laser endomicroscopy (pCLE) in patients with barrett's esophagus. Gastroenterology (DDW 2011) (2011)
11. Barnett, V.: Sample Survey principles and methods. Hodder Arnold (1991)
12. Meng, X.L., Rosenthal, R., Rubin, D.B.: Comparing correlated correlation coefficients. Psychological Bulletin **111**(1) (1992) 172–175