Semi-automated Query Construction for Content-Based Endomicroscopy Video Retrieval

Marzieh Kohandani Tafreshi[1,2,⋆], Nicolas Linard[2,⋆], Barbara André[2], Nicholas Ayache[1], and Tom Vercauteren[2]

[1] Inria Asclepios Project-Team, Sophia Antipolis, France
[2] Mauna Kea Technologies, Paris, France

Abstract. Content-based video retrieval has shown promising results to help physicians in their interpretation of medical videos in general and endomicroscopic ones in particular. Defining a relevant query for CBVR can however be a complex and time-consuming task for non-expert and even expert users. Indeed, uncut endomicroscopy videos may very well contain images corresponding to a variety of different tissue types. Using such uncut videos as queries may lead to drastic performance degradations for the system. In this study, we propose a semi-automated methodology that allows the physician to create meaningful and relevant queries in a simple and efficient manner. We believe that this will lead to more reproducible and more consistent results. The validation of our method is divided into two approaches. The first one is an indirect validation based on per video classification results with histopathological ground-truth. The second one is more direct and relies on perceived inter-video visual similarity ground-truth. We demonstrate that our proposed method significantly outperforms the approach with uncut videos and approaches the performance of a tedious manual query construction by an expert. Finally, we show that the similarity perceived between videos by experts is significantly correlated with the inter-video similarity distance computed by our retrieval system.

1 Introduction

Probe-based Confocal Laser Endomicroscopy (pCLE) enables the endoscopist to acquire real-time *in situ* and *in vivo* microscopic images of the epithelium during an endoscopy. As shown in [3], content-based retrieval (CBR) methods may provide interpretation support for the endoscopist, helping him or her in making an informed decision and establishing a more accurate pCLE diagnosis. However, the selection of adapted query can be quite challenging and time-consuming for the user of such a CBR system. Also, because of the complexity of such manual query construction, the CBR system may not have a sufficient reproducibly and may be subject to large intra and inter-observer variability.

The approach presented in this paper allows physicians to efficiently create reproducible queries in a semi-automated fashion. This allows to boost retrieval

⋆ Authors have contributed equally to the paper.

P. Golland et al. (Eds.): MICCAI 2014, Part I, LNCS 8673, pp. 89–96, 2014.

performance when compared to using uncut videos as queries. It also allows us to approach the performance of carefully constructed queries by an expert. To achieve this, our query construction approach is decomposed in two steps.

In the first step, we perform an automated temporal segmentation of the original video into a set of subsequences of interest. The segmentation is based on kinematic stability assessment. Since endomicroscopy is a handheld interventional modality, users often swipe a region of interest to look for diagnostically relevant criteria. In this work we leverage the observation that spatial stability across time is related to the informativeness of the images to design a first video stream temporal segmentation algorithm dedicated to endomicroscopy.

The second step consists in a fast user selection of a subset of the segmented subsequences. The physician is simply asked to keep or discard the subsequences provided by the first step. Although each of the possible subsequences may still contain images of different tissue type, the segmentation step makes each subsequence much more self-consistent than the original uncut video.

Once a query has been constructed, our method relies on the video CBR method presented in [2]. This system is based on the Bag-of-Visual-Words (BoW), a review of which can be found in [15]. Instead of relying on salient features, [2] uses a regular grid of descriptors at a fixed scale to construct a visual signature. This signature is then used to index and query a database of annotated cases.

Evaluation of CBR systems is known to be a difficult task. In our work, similarly to [3], two validation methodologies with different strengths are used. The first indirect one uses the retrieval results and a k-nearest neighbors (k-NN) voting scheme to classify each video. This approach benefits from the fact that, in most clinically validated databases, each video is associated with a histopathologically validated diagnosis. The k-NN classification is more a quantitative evaluation and serves only as a CBR evaluation proxy. The second validation methodology compares the inter-video distances computed by the CBR method with the perceived visual similarities experienced by experts.

2 A Temporal Segmentation with User Selection Pipeline

As illustrated in Fig. 1, our approach to query construction and CBR works as follows. During a procedure, the physician acquires pCLE videos in real time. The acquired frames are stored in a bounded circular FIFO buffer. At any moment during the intervention the user may want to consult the annotated database to provide him or her with visually similar cases that have been confirmed by histopathology examination. At this point, the acquisition is paused and our software displays the image buffer with a timeline that shows the automatic temporal segmentation. The user is then asked to briefly review each segmented subsequence and click on the ones that are of interest to him. This simplified interaction allows the user to construct a fast and reproducible query with sufficient visual similarity within and between the selected subsequences. Because all this happens during the procedure, our temporal segmentation needs to be running in real-time. Then, the user-chosen subset of subsequences is used to

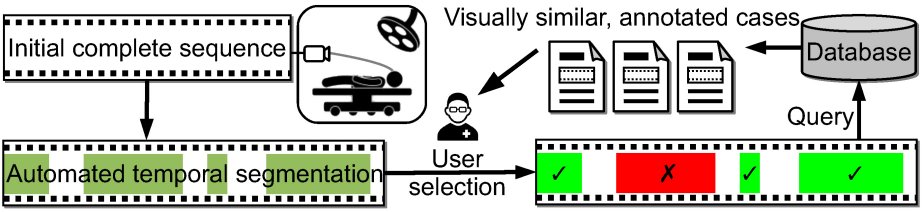

Fig. 1. An overview of our semi-automated query construction algorithm as used in a clinically relevant CBR pipeline

create a single visual signature to query the database. The most visually similar cases are presented to the physician along with their annotations.

3 Temporal Segmentation from Kinematic Stability

As outlined in [6,7], temporal video segmentation is a key step in most existing video management tools. Many different types of algorithms have been developed to perform temporal video segmentation. Early techniques focused on cut-boundary detection or image grouping using pixel differences, histogram comparisons, edge differences, motion analysis and the like, while more recent methods such as presented in [5] have also used image similarity metrics, classification and clustering to achieve the same goal. In some applications as in [11,12], the problem of temporal video segmentation may be reformulated as a classification problem that distinguishes between informative and noise images.

Our approach in this work relies on the observation that, during an pCLE procedure, the user will navigate the imaging probe across the region of interest and will typically stay longer and remain more stable onto areas that catches his or her interest. As such, in our application, kinematic stability may serve as a proxy to characterize the interest of an image within a sequence.

Image registration-based approaches can be used to identify kinematically stable temporal regions. This can be done by actually registering temporally consecutive images and then analyzing the quality of the spatial transformation. For exemple, [13] relies on real-time registration algorithms. Kinematic stability assessment may also be done by using only a subset of the steps of an image registration algorithm and analyze the quality of the results provided by this subset. In this paper, feature matches are analyzed in terms of *local* spatial consistency so as to obtain a result that is more robust to modeling error and to tissue deformation than looking for an accurate spatial transformation.

In [2], the authors have shown that, although the typical feature detectors from computer vision described in [15,8] are not suitable for endomicroscopy, one may rely on a regular grid of Scale Invariant Feature Transform (SIFT) descriptors for the purpose of defining visual signatures. Because our method requires features both to assess kinematic stability and to create visual signatures, we propose to rely on the same set of SIFT descriptors so as to reduce the computational

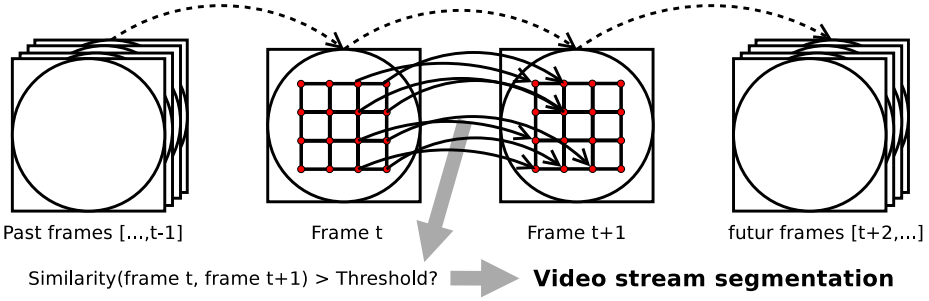

Fig. 2. Our temporal segmentation algorithm based on kinematic stability assessment

requirements. In the field of computer vision, regular grids of descriptors have also recently been used for image matching and video analysis with compelling results [9,14]. However, to the best of our knowledge, our work is the first to rely on such regular grids of descriptors to evaluate kinematic stability by looking at the local consistency of feature matches.

As illustrated in Fig. 2, our method starts by decomposing each frame into a grid of SIFT descriptors and matching the descriptors between consecutive frames. Local consistency of the matches is assessed by making each match vote for a translation in a vote map. A kinematic stability criterion is then computed from the vote map and a threshold on this criterion is used to discriminate kinematically stable and unstable frame transitions.

In more detail, we associate a grid point at time t with one at $t+1$ by minimizing the Euclidean distance between the corresponding descriptors. Although other distances may be used, Euclidean distance allows for the use of efficient approximate nearest neighbor algorithms. Similarly to [10], we filter out bad matches by comparing, for each source grid point, the best match with subsequent ones. When working with sparsely located features, one may simply compute the ratio of the distances for the first and second best match and use a threshold on this ratio. However, in the case of a regular grid, the regions covered by adjacent descriptors may show a large overlap. This implies that the first and second best match will often be very close in terms of descriptors distance. We therefore propose to compute the distance ratio between the first and n^{th} best matches. We choose n such that the description regions associated with a grid point and with its n^{th} spatial neighbor grid point have no overlap.

From the set of filtered matches, our goal is not to find an accurate spatial transformation as in [9], but to define a computationally efficient kinematic stability criterion. For this purpose, and although we know that pCLE videos suffer from motion distortions and tissue deformation, we rely on a local translation model that was proven to work in [13]. As such, we build a map where each match votes for its translation. It should however be noted that not all translations can receive the same maximum number of votes. To account for this potential bias, the vote map is weighted according to the maximum number of potential voters

per voting bin. This normalization is computed by the autocorrelation of a mask image that represent the spatial organization of the description grid.

For a simple translation across two frames, we observe a single main peak in the normalized vote map. However, in the standard case of more complex transformations, we may observe several peaks, blobs or ridges in the vote map, all of them corresponding to locally consistent translations. To account for this effect, we define our kinematic stability criterion by adding up all votes that are above a predefined consistency threshold. Finally, a frame transition is considered kinematically stable if the kinematic stability criterion is above a predefined stability threshold. This stability threshold was obtained by optimizing the correlation between automatic and manual segmentation results.

4 Video Retrieval Based on Bag of Visual Words

Given a temporal segmentation and a user selection within the subsequences, our aim now is to query a database with the selected video parts. In our study we rely on the approach of [2] but do not use their mosaicing strategy for computational reasons. The Bag of Visual Words method is adapted to pCLE retrieval by working with a regular grid of SIFT descriptors at a fixed scale. Avoiding scale invariance is a requirement for pcLE videos, where for example, in colonic polyps, a mesoscopic crypt and a microscopic goblet cell have both a rounded shape, but have different sizes. From the sets of SIFT descriptors, each image within the selected video part gets associated with a visual signature. By averaging these signatures, each subsequence and each video are associated with a visual signature that can be used for retrieval purposes.

5 Results: Classification and Perceived Similarity

To evaluate the relevance of our retrieval results, two procedures are used: an indirect one based on classification and a direct one based on perceived similarity. In both cases, we rely on a pCLE database of colonic polyps sequences that were retrospectively collected from pCLE procedures performed in the Mayo Clinic in Jacksonville, Florida, USA. This database is composed of 118 pCLE videos (35 benign, 83 neoplastic) that were acquire from 66 patients [4]. The length of these videos ranges from 1 second to 4 minutes and their median duration is 28.2 secs. Long videos may contain different tissues, however as such most videos are sufficiently short to display a single tissue type. The parameters of the retrieval method we used are the one provided in [2].

For the classification evaluation, a straightforward k-NN classification is performed and its accuracy is estimated. Two classes are considered, benign and neoplastic. For these videos, the pCLE diagnosis is matched to the *gold standard* established by a pathologist after the histological review of biopsies acquired on the pCLE imaging spots. Given the small number of videos contained in the database, each of the videos is used for both training and testing. A leave-one-patient-out (LOPO) cross-validation allows us to respect the independence

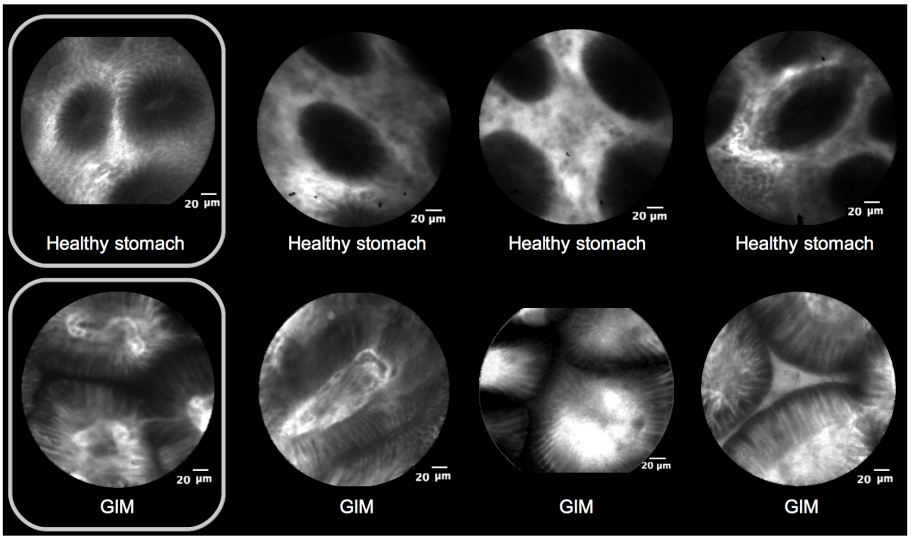

Fig. 3. Retrieval examples of our colonic polyps datasets. The query video is on the left and its 3 most similar videos represented by single frames are on the right. B. indicates Benign and N. Neoplastic.

assumption that is challenged by the fact that several videos are acquired from the same patient. Although an indirect evaluation, these experiments allow for quantitative evaluation of our method and rely on an objective ground truth.

For the perceived similarity experiments, we rely on the database of [1]. Using an online survey tool, a pairwise similarity ground-truth between pCLE videos was estimated by 17 human observers, ranging from middle expert to expert in establishing pCLE diagnosis, who are completely ignorant to the video metadata such as the pCLE diagnosis. Each video couple gets assigned a score from the following four-points Likert scale: *very dissimilar*, *rather dissimilar*, *rather similar* and *very similar*. In total 4,836 similarity scores were given for 2,178 distinct video couples. Thus 16,2% of all 13,343 distinct video couples were scored. We can then compare the visual similarity distance computed by our method with the perceived one. These results shed a different light on the evaluation. They reflect our target application better but rely on a more subjective ground truth with high inter-observer variability.

In this paper, four different methods are compared. The first one is that of [3], without the mosaicing part, in which an expert carefully constructs the queries. The second one relies on uncut videos. The third one uses the entire set of subsequences generated by the automated temporal segmentation algorithm. Finally, the fourth one is our proposed semi-automated method. For each compared method, the same number of sequences is used. To enable a fair comparison between the methods, the following procedure was used for the selection of the subsequences in the semi-automated approach. Instead of asking an expert to

Table 1. Evaluation of the performance of our proposed semi-automated approach in comparison to state of the art methods. The evaluation is performed both indirectly in terms of classification and directly in terms of correlation with perceived similarity.

	Classification results			Perception
Method	**Accuracy**	**Sensitivity**	**Specificity**	**Spearman** ρ
Complete uncut sequences	72.9 %	72.3 %	74.3 %	35.9 %
Fine expert temporal segmentation	94.1 %	96.4 %	88.6 %	52.8 %
Automated temporal segmentation	61.9 %	56.6 %	74.9 %	31.6 %
Proposed semi-automated method	89.9 %	90.4 %	88.6 %	48.8 %

perform the selection, we re-used the careful query construction from the first method. For each temporal segment, if it contains at least one frame that was chosen in the careful query construction, the temporal segment is marked as selected. This allows for an unbiased comparison of both methods.

In all cases, to ensure strong self-consistency within the annotated database, only the carefully constructed sequences by the expert were used in the training phases. For each step, different experts from different clinical trials were consulted. Particularly, 17 observers participated to create the visual similarity ground-truth using the online VSS tool. We believe that the fact that these steps are performed by different experts leads to unbiased results.

As shown in Table 1, our semi-automated method significantly outperforms the two automated ones and approaches to performance of the reference manual one. The accuracies with uncut sequences and automated temporal segmentation are indeed low, but even with these baselines, the correlation between the computed and the perceived similarity is higher than chance with statistical significance. A McNemar's test show that, with statistical significance, our proposed semi-automated method is better than the uncut sequences method (p-value < 0.0005 for $k = 10$) and than the automated temporal segmentation method (p-value $< 10^{-6}$ for $k = 10$) in terms of classification. We also observe that the difference between our method and the fine expert temporal segmentation method is not statistically significant.For the correlation with perceived similarity, we performed a Steiger's Z-test applied to the Spearman ρ correlation coefficient. This test indicates that the improvement of our method over the uncut sequences method and over the automated temporal segmentation method is statistically significant (p-value $< 10^{-6}$). Unsurprising, the fine expert temporal segmentation statistically outperforms our method.

6 Discussion

This study proposes a fast and semi-automated approach to constitute a relevant and informative query to submit to the retrieval system. Our results have

demonstrated that the classification results and the perceived similarity remain consistent in comparison with the results obtained using queries which are manually selected by an expert. Future work will improve the selection of the query, and aim at achieving better fully automated query construction. Nonetheless, we believe that the proposed methodology makes content-based retrieval techniques closer to clinical utility.

References

1. André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: Retrieval evaluation and distance learning from perceived similarity between endomicroscopy videos, 289–296 (2011)
2. André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: A smart atlas for endomicroscopy using automated video retrieval 15(4), 460–476 (2011)
3. André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: Learning semantic and visual similarity for endomicroscopy video retrieval 31(6), 1276–1288 (2012)
4. 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.C., Wallace, M.B.: Comparison of probe-based confocal laser endomicroscopy with virtual chromoendoscopy for classification of colon polyps 138(3), 834–842 (2010)
5. Cooper, M., Liu, T., Rieffel, E.: Video segmentation via temporal pattern classification 9(3), 610–618 (2007)
6. Gargi, U., Kasturi, R., Strayer, S.H.: Performance characterization of video-shot-change detection methods 10(1), 1–13 (2000)
7. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A survey on visual content-based video indexing and retrieval 41(6), 797–819 (2011)
8. Li, B., Meng, M.H.: Capsule endoscopy video boundary detection, 373–78 (June 2011)
9. Liu, C., Yuen, J., Torralba, A.: SIFT flow: Dense correspondence across scenes and its applications 33(5), 978–994 (2011)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints 60, 91–110 (2004)
11. Oh, J., Hwang, S., Lee, J., Tavanapong, W., Wong, J., de Groen, P.C.: Informative frame classification for endoscopy video 11(2), 110–127 (2007)
12. Sun, Z., Li, B., Zhou, R., Zheng, H., Meng, M.H.: Removal of non-informative frames for wireless capsule endoscopy video segmentation, 294–299 (August 2012)
13. Vercauteren, T., Meining, A., Lacombe, F., Perchant, A.: Real time autonomous video image registration for endomicroscopy: Fighting the compromises. In: Conchello, J.A., Cogswell, C.J., Wilson, T. (eds.) Proc. SPIE BIOS - Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XV. SPIE, San Jose (2008)
14. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition 103(1), 60–79 (2013)
15. Zhang, J., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study 73, 213–238 (2007)