

# Unsupervised Feature Learning for Endomicroscopy Image Retrieval

Yun Gu<sup>1,2,3</sup>, Khushi Vyas<sup>3</sup>, Jie Yang<sup>1,2,\*</sup>, and Guang-Zhong Yang<sup>3</sup>

<sup>1</sup> School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, CHINA

<sup>2</sup> Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, CHINA

<sup>3</sup> Hamlyn Centre for Robotic Surgery, Imperial College London, London, UK  
(\*Jie Yang is the corresponding author.)

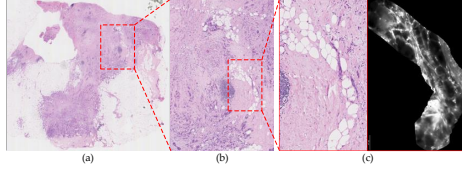
**Abstract.** Learning the visual representation for medical images is a critical task in computer-aided diagnosis. In this paper, we propose Unsupervised Multimodal Graph Mining (UMGM) to learn the discriminative features for probe-based confocal laser endomicroscopy (pCLE) mosaics of breast tissue. We build a multiscale multimodal graph based on both pCLE mosaics and histology images. The positive pairs are mined via cycle consistency and the negative pairs are extracted based on geodesic distance. Given the positive and negative pairs, the latent feature space is discovered by reconstructing the similarity between pCLE and histology images. Experiments on a database with 700 pCLE mosaics demonstrate that the proposed method outperforms previous works on pCLE feature learning. Specially, the top-1 accuracy in an eight-class retrieval task is 0.659 which leads to 10% improvement compared with the state-of-the-art method.

**Keywords:** Probe-based Laser Endomicroscopy, Histology, Graph Mining, Feature Learning

## 1 Introduction

Probe based Confocal laser endomicroscopy (pCLE) is a popular optical biopsy technique capable of in situ and in-vivo imaging of tissue architecture with microscopic resolution. Using flexible fiber bundle and miniaturized optics, it provides clinicians with real-time access to histological information during surgical procedures and has demonstrated promising sensitivities and specificities in various preclinical and clinical studies, including in the gastro-intestinal tract, urinary tract, breast and respiratory system.

In surgical procedures, histopathological examination of biopsy samples by trained pathologists is the gold standard for disease diagnosis, grading and classification. Although pCLE enables to acquire in vivo microscopic images that resemble closely to histology images, in vivo diagnosis is still challenging for many clinicians who have little histopathology expertise and training. Further



**Fig. 1.** The region correspondence between histology images and pCLE mosaics. (a) is the original histology image with the high resolution  $86k \times 68k$ ; (b) is a subregion of (a) with 40x zoom in; (c) is a subregion of (b) within the field-of-view (FoV) of endomicroscopy and the corresponded pCLE mosaic.

the high variability in the appearance of pCLE images and presence of atypical conditions makes it difficult to provide accurate diagnosis by manual identification.

In the past few years, Content-based medical image retrieval (CBMIR) technique has been increasingly applied for accurate clinical diagnosis and assessment of medical images [1, 2]. Given the query image, CBMIR systems return the most related examples from the large databases which could provide informative diagnosis/decision support to clinicians. The existing literature on CBMIR for pCLE images mainly focuses on learning discriminative visual features to improve the retrieval accuracy which can be divided into two categorizations including *unimodal methods* [3, 4] and *multimodal methods* [5]. For unimodal methods, the discriminative features are learned based on pCLE images only. Examples of unimodal methods include densely sampled SIFT[4] and the metric learning based on SIFT [3] for CBIR tasks with pCLE videos. Although previous works demonstrate promising performance, the pCLE systems have the limited field of view, particularly when compared to histology slides which means that only a small number of morphological features can be visualized in each image. As a result, discriminating between benign, atypical and neoplastic lesions solely based on pCLE data might be a difficult task. On the other hand, a multimodal framework which includes both pCLE and histology images can bridge this gap and enhance the discriminative feature of pCLE for accurate decision support. Recently, Gu et.al [5] demonstrated that the multimodal framework contributes to promising accuracy in pCLE mosaics classification.

Although the multimodal method in [5] aids to enhance the discriminative features by using pCLE and histology images, a one-to-one correspondence is required to achieve the desired accuracies. As shown in Fig 1, due to huge differences in the field of view, the pCLE mosaic only corresponds to a small region in histology image. Although we can get an approximate location during the scanning, it is still time-consuming to find the corresponded region from histology images for a specific mosaic. Particularly for freehand pCLE imaging, manually finding a large number of pCLE-histology pairs to learn the discriminative features could be challenging.

In this paper, we propose to overcome this challenge by developing an Un-supervised Multimodal Graph Mining (UMGM) framework to learn discrimina-

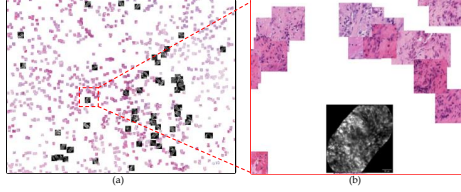
tive features for pCLE mosaics. Our approach is inspired by recent advances in metric learning based on graph analysis[6, 7]. The first step of our work is to extract similar and dissimilar histology patches for a specific pCLE mosaic without supervised information. (In this paper, we use 'similar/dissimilar' and 'positive/negative' interchangeably.) In order to extract sufficient data pairs, we tend to discover the latent similarity between pCLE mosaics and histology patches within the field-of-view (FoV) as well as the region out of FoV. Although the pCLE mosaic corresponds to a small FoV in the complete histology image, we observe that there are large amount of regions which are out of FoV with similar tissue morphology of specific region scanned by pCLE. These regions can cover more variance of the structure of cells that tend to provide extra information for feature learning. Inspired by this observation, a multiscale multimodal graph is built based on pCLE mosaics and histology patches. The latent similarity among pCLE and histology patches is mined by leveraging graph-based analysis over a large collection of histology patches. Specifically, we propose to use a cycle consistency criterion for mining positive pairs and the geodetic distance to find the hard negative pairs. The discriminative feature of pCLE is learned via minimizing the distance between positive pairs and maximizing the distance between negative pairs. We validate the performance of the representation on dataset with 45 cases which consists of 700 pCLE mosaics and their corresponding histology slides. The experiments demonstrate that the proposed method outperforms previous single modal approach [4] and the supervised multimodal method [5].

## 2 Methodology

### 2.1 Building and Mining the Multimodal Graph

In this paper, pCLE mosaics are denoted by  $\{X_i^P\}, i = 1, \dots, n_P$  where  $n_P$  is the number of pCLE mosaics. Each mosaic is matched with a small FoV in the whole histology image which are denoted by  $\{Y_i^P\}, i = 1, \dots, n_P$ . We randomly select patches  $Y_i^{R,s}$  with size  $s$  from histology images where  $s \in \{128, 256, 512, 1024\}$ . For pCLE mosaics and histology patches, dense Scale Invariant Feature Transformation (dense-SIFT) [4] is adopted as visual representation. In order to learn the discriminative features, we build the multimodal graph  $G = \{V, E\}$ . The nodes  $V$  in the graph are composed with patch nodes  $V^p$  and anchor nodes  $V^a$ . The patch nodes are histology patches randomly extracted from the whole histology images. The elements of anchor nodes  $V^a$  in the graph are doublets of pCLE mosaics and their corresponding histology patches where  $V^a = \{(X_i^P, Y_i^P), i = 1, \dots, N_p\}$ . The directed edges  $E = \{e_{i,j}\}$  indicate the connection between nodes where  $e_{i,j} = 1$  if  $V_j$  belongs to the k-nearest neighbors of  $V_i$ . The edge weight  $w_{i,j}$  is defined between nodes  $V_i$  and  $V_j$  is defined as follows:

$$w_{i,j} = \begin{cases} \|Y_i^R - Y_j^R\|_2, & \text{if } V_i \in V^p \wedge V_j \in V^p \\ \|Y_i^R - Y_j^P\|_2, & \text{if } V_i \in V^p \wedge V_j \in V^a \\ \|Y_i^P - Y_j^R\|_2, & \text{if } V_i \in V^a \wedge V_j \in V^p \end{cases}$$



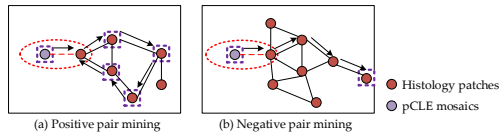
**Fig. 2.** The visualization of graph based on tSNE[8]. (a) is the overview of the multimodal graph; (b) is an example of anchor node and its nearest neighbours

where  $\|\cdot\|_2$  is the  $L_2$  norm. The distance between patch nodes are measured by the Euclidean distance based on their visual feature. Since the similarity between pCLE mosaics and histology patches cannot be directly calculated yet, the distance between anchor nodes and patch nodes is determined by the distance between the histology images only. The visualization of the graph via tSNE[8] is shown in Fig. 2. The anchor node and its nearest neighbours in Fig. 2(b) represent the dense scatter of cell nucleus.

Based on the graph, we tend to extract positive pairs and negative pairs for feature learning. Starting from anchor node  $V_i^a$ , we define that patch node  $V_j^p$  is an  $n$ -order  $k$ -nearest neighbor of the node  $V_i^a$  if there exists a directed path of length  $n$  from node  $V_i^a$  to node  $V_j^p$ . As defined in [6], if  $V_i$  belongs to its own  $n$ -order  $k$ -nearest neighbors, we can obtain a directed cycle as follows:

$$V_i \in \mathcal{N}_k^{(n)}(V_i), n \geq 2 \quad (1)$$

where  $\mathcal{N}_k^{(n)}(V_i)$  is the set of  $n$ -order  $k$ -nearest neighbors for  $V_i$ . As shown in Fig. 3(a), for each anchor node in multimodal graph, we search its  $n$ -order  $k$ -nearest neighbours and detect its  $n$ -order cycle. An  $n$ -order cycle contains the histology image in anchor node and  $n-1$  histology patches which can finally generate  $n$  different pairs of pCLE and histology matching. We combine these positive samples with the pCLE mosaic in the anchor node, which are denoted by  $\{X_i^{(pos)}, Y_i^{(pos)}\}$ , as similar pairs which are used for feature learning. Compared with conventional nearest-neighbour schemes for positive sample mining, the nodes in a cycle indicate the consistent relationship and robustness to outliers. Besides the positive pairs, we also tend to extract negative pairs which are dissimilar samples to enhance the visual representation. In this paper, we use the geodesic distance between anchor nodes and patch nodes in the multimodal graph. As shown in Fig. 3(b), we first use the Floyd-Warshall algorithm [9] to find the shortest paths between each node in the graph. The geodesic distance  $g_{ij}$  is the accumulated edge weights along the shortest path from anchor node  $V_i^a$  to patch node  $V_j^p$ . We then perform random selection among those image pairs with the geodesic distance larger than the threshold  $d_m$  given as negative samples which are denoted by  $\{X_i^{(neg)}, Y_i^{(neg)}\}$ .



**Fig. 3.** Unsupervised graph mining for positive/negative pair extraction.

## 2.2 Discriminative Feature Learning

Based on the positive pairs  $\{X_i^{(pos)}, Y_i^{(pos)}\}$  and negative pairs  $\{X_i^{(neg)}, Y_i^{(neg)}\}$ , we tend to learn two transformations  $f_X$  and  $f_Y$  which independently map the pCLE mosaics and histology images to a latent feature space. These transformations should promise that the positive pairs are similar in the latent space while the negative pairs are not similar. Therefore, we build the loss function penalizing the distances of positive pairs greater than a threshold, and distances of negative pairs smaller than the same threshold. The objective function is illustrated as follows:

$$\min_{(X_i, Y_i) \in \{X_i^{(pos)}, Y_i^{(pos)}\}} l(\alpha_i^{(pos)} \|f_X(X_i) - f_Y(Y_i)\|^2 - 1) - \lambda \sum_{(X_i, Y_i) \in \{X_i^{(neg)}, Y_i^{(neg)}\}} l(\alpha_i^{(neg)} \|f_X(X_i) - f_Y(Y_i)\|^2 - 1) \quad (2)$$

where  $l(\cdot)$  is the generalized logistic loss function  $l(x) = \log(1 + e^{\beta x}) / \beta$ .  $\alpha_i^{(pos)}$  and  $\alpha_i^{(neg)}$  are sample weights for positive and negative pairs respectively. The sample weights are determined by the accumulated distance from anchor nodes to patch nodes. Although many different kinds of functions can be used to define  $f_X$  and  $f_Y$ , we adopt the commonly used linear function form where  $f_X(x) = W_X^T x$  and  $f_Y(y) = W_Y^T y$ . The problem in Eq.(2) can be solved via simple gradient descent approaches as used in [10]. For pCLE mosaics, the discriminative representation can be finally obtained by multiplying the transformation matrix  $W_X^T X$ .

## 3 Experiments

### 3.1 Dataset and Experimental Settings

The dataset is collected by a pre-clinical pCLE system (Cellvizio, Mauna Kea Technologies, Paris, France) as described in [11]. Breast tissue samples are obtained from 45 patients that are diagnosed with three main classes including normal, benign and and neoplastic. Eight sub-classes are defined based on tissue and diagnosis information. For normal cases, the mosaics are further classified into adipose tissue , elastic fibres , collagen fibres and normal breast lobule. For benign cases, the subclasses contain Dilated Breast Ducts and Fibroadenoma. For neoplastic cases, diagnosis result supports the existence of specific lesion including ductal carcinoma in situ (DCIS) and invasive cancers. After completion

of pCLE imaging, each sample underwent routine histopathology processing to generate the histology images. We finally obtain 700 pCLE mosaics and 144 of them are matched with histology patches. In training procedure, we use 144 pairs of pCLE mosaics and histology images as anchor nodes while the rest of data are used in testing phase. The complete histology image from all 45 patients are scanned with multi-scale windows and 8,000 patches are randomly selected to build the multimodal graph. For visual representation, we generate 500D BoW SIFT feature for mosaics and 200D BoW SIFT feature for histology images. The dimension of the latent space learned by the proposed feature is set to 64.

In order to evaluate the proposed method, we conduct the retrieval tasks based on both main-classes and sub-classes. The top-1 and top-5 accuracy are both reported according to the retrieved result. Several baselines are implemented in this paper for comparison including dense-SIFT in [4] and MVMME in [5]. The proposed method with only positive pair mining is also set as baseline for comparison. All experiments are performed by 10-fold cross-validation. The hardware platform for evaluation is a PC with Intel 2.4GHz CPU and 16GB RAM. Methods are implemented with MATLAB.

In the experiments, we set the number of nearest neighbours to  $k = 5$ . The length of cycle in Eq.(1) is set to 5. After the positive and negative mining, we finally obtain 720 pairs of similar examples and 5000 dissimilar samples for feature learning.

### 3.2 Results

In this section, we present the numerical results of retrieval tasks on pCLE dataset. Table 1 shows the retrieval performance of multiple baseline approaches and the proposed UMG. The following observations can be derived:

- Compared with approach based on pCLE mosaics only, the accuracy of retrieval system is effectively improved by multimodal embedding strategies. The corresponding histology images of pCLE mosaics can provide informative features to distinguish different types of tissues.
- In retrieval tasks for three main-classes, we can only achieve slight improvement with 2% higher accuracy than unimodal approaches. Since the main-class task combines all fine-grained tissues into the same class, the differences among sub-classes are ignored. Even with unimodal approaches, the performance is still promising.
- In retrieval tasks for eight sub-classes, the improvement gained from the proposed method is significant. Compared with previous works, the top-1 retrieval accuracy is 0.659 while DenseSIFT and MVMME are below 0.6. In order to demonstrate in details, we take DCIS and Invasive as example. The characteristic of DCIS is the thickened ductal epithelium. However, in some cases, the duct structure will not be completely scanned while the rest tissues within the FoV are scattered nucleus. The scattered nucleus can also appear in invasive carcinomas which lead to incorrect retrieval result. In our approach, the cycle the tissues from the whole histology slide can provide a

**Table 1.** Performance of CBMIR tasks. UMGM-Pos indicates the proposed method with only positive examples

Method	Main-Class Top1	Main-Class Top5	Sub-Class Top1	Sub-Class Top5
DenseSIFT	0.875	0.937	0.562	0.818
MVMME	0.886	0.943	0.568	0.829
UMGM-Pos	0.891	0.937	0.619	0.835
UMGM	<b>0.892</b>	<b>0.962</b>	<b>0.659</b>	<b>0.864</b>
Method	DCIS Top1	DCIS Top5	Invasive Top1	Invasive Top5
DenseSIFT	0.244	0.511	0.648	0.956
MVMME	0.267	0.467	0.713	0.978
UMGM-Pos	0.333	0.533	0.692	0.989
UMGM	<b>0.378</b>	<b>0.578</b>	<b>0.747</b>	<b>0.989</b>

better view of specific structure. As shown in Table 1, the proposed method effectively improve the accuracy for DCIS and Invasive tissues.

- Compared with the graph mining based on only positive sample selection, the negative samples can lead to further improvement on retrieval accuracy.

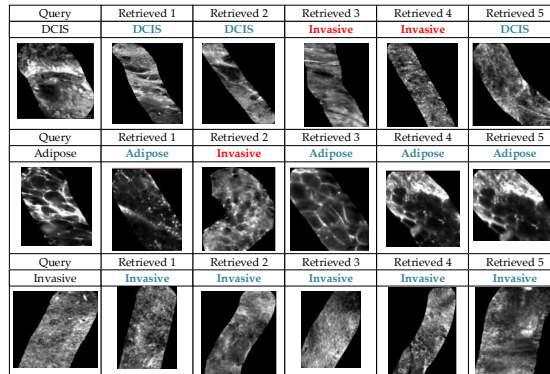
We also present some examples of the retrieval based on the proposed method in Fig. 4 with three typical sub-classes. The texts in blue indicate the true response while the texts in red refer to the false result. It can be observed that the proposed method can return proper examples according to the query in many cases with both promising accuracy and good diversity. Some failure cases are also presented in Fig. 4.

## 4 Conclusion

In this paper, we propose an Unsupervised Multimodal Graph Mining (UMGM) framework to learn discriminative features for endomicroscopy retrieval. A multiscale multimodal graph is built based on pCLE mosaics and histology patches. The latent similarity among pCLE and histology patches is mined by leveraging graph-based analysis over a large collection of histology patches. The discriminative feature of pCLE is learned via minimizing the distance between positive pairs and maximizing the distance between negative pairs. Compared with the previous works, the embedding of multimodal images contributes to higher accuracy on retrieval tasks without supervised information.

## 5 Acknowledgement

This work is partially supported by NSFC, China (No:61572315,6151101179) and 973 Plan, China (No. 2015CB856004). Yun Gu is also supported by Chinese Scholarship Council (CSC). The tissue specimens were obtained from consented patients using the Imperial tissue bank ethical protocol following the R-12047 project.



**Fig. 4.** Examples of retrieval tasks.

## References

- Jiang, M., Zhang, S., Huang, J., Yang, L., Metaxas, D.N.: Scalable histopathological image analysis via supervised hashing with multiple features. *Medical Image Analysis* **34** (2016) 3–12
- Zhang, X., Su, H., Yang, L., Zhang, S.: Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval. In: *CVPR*. (2015) 5361–5368
- André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: A smart atlas for endomicroscopy using automated video retrieval. *Medical image analysis* **15**(4) (2011) 460–476
- André, B., Vercauteren, T., Perchant, A., Buchner, A.M., Wallace, M.B., Ayache, N.: Endomicroscopic image retrieval and classification using invariant visual features. In: *IEEE ISBI 2009, IEEE* (2009) 346–349
- Gu, Y., Yang, J., Yang, G.Z.: Multi-view multi-modal feature embedding for endomicroscopy mosaic classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. (2016) 11–19
- Li, D., Hung, W.C., Huang, J.B., Wang, S., Ahuja, N., Yang, M.H.: Unsupervised visual representation learning by graph-based consistent constraints. In: *ECCV*. (2016)
- Zhai, X., Peng, Y., Xiao, J.: Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: *AAAI’13*. (2013) 1198–1204
- Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. *JMLR* **9**(Nov) (2008) 2579–2605
- Floyd, R.W.: Algorithm 97: shortest path. *Communications of the ACM* **5**(6) (1962) 345
- Mignon, A., Jurie, F.: Cmml: a new metric learning approach for cross modal matching. In: *Asian Conference on Computer Vision*. (2012) 14–pages
- Chang, T.P., Leff, D.R., Shousha, S., Hadjiminias, D.J., Ramakrishnan, R., Hughes, M.R., Yang, G.Z., Darzi, A.: Imaging breast cancer morphology using probe-based confocal laser endomicroscopy: towards a real-time intraoperative imaging tool for cavity scanning. *Breast cancer research and treatment* **153**(2) (2015) 299–310