Multi-View Multi-Modal Feature Embedding for Endomicroscopy Mosaic Classification

Yun Gu
School of Biomedical Engineering
Shanghai Jiao Tong University
geron762@sjtu.edu.cn

Jie Yang
School of Biomedical Engineering
Shanghai Jiao Tong University
jieyang@sjtu.edu.cn

Guang-Zhong Yang
Hamlyn Centre for Robotic Surgery
Imperial College of London
g.z.yang@imperial.ac.uk
Please contact: jieyang@sjtu.edu.cn

Abstract

Probe-based confocal laser endomicroscopy (pCLE) is an emerging tool for epithelial cancer diagnosis, which enables in vivo microscopic imaging during endoscopic procedures. As a new technique, definite clinical diagnosis is still referenced to the gold standard histology images. In this paper, we propose a Multi-View Multi-Modal Embedding framework (MVMME) to learn representative features for pCLE videos exploiting both pCLE mosaic and histology images. Each pCLE mosaic is represented by multiple feature representations including SIFT, Texton and HoG. A latent space is discovered by embedding the visual features from both mosaics and histology images in a supervised scheme. The features extracted from the latent spaces can make use of multi-modal imaging sources that are more discriminative than unimodal features from mosaics alone. The experiments based on real pCLE datasets demonstrate that our approach outperforms, with statistical significance, several single-view or single-modal methods. A binary classification accuracy of 96% has been achieved.

1. Introduction

Probe-based Confocal Laser Endomicroscopy (pCLE) enables endoscopist to acquire real-time *in situ* and *in vivo* microscopic images of the epithelium during an endoscopy procedure. As mentioned in [3], the main task for the endoscopists is to establish a diagnosis from the acquired pCLE videos, by relating a given appearance of the epithelium to a specific pathology. Due to the processing involved, tissue characterization is often performed offline in current prac-

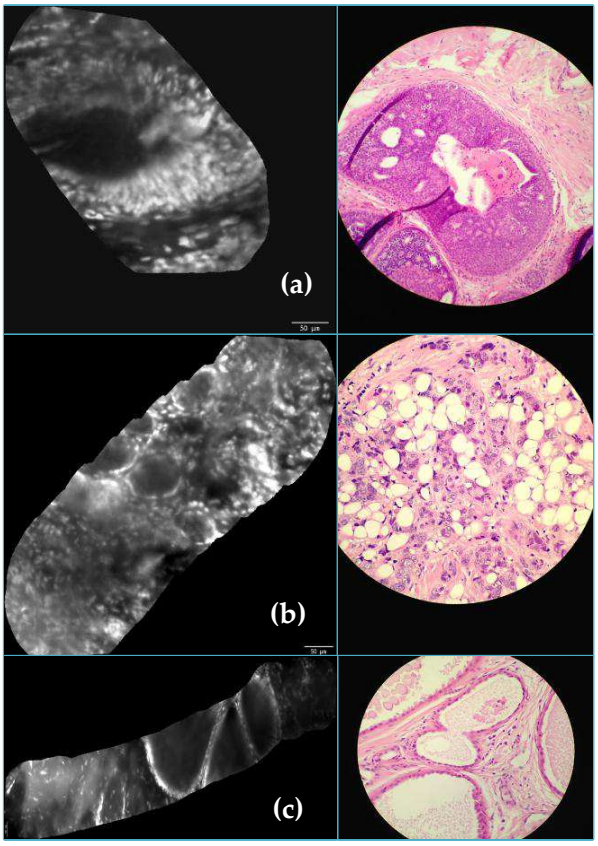


Figure 1. Morphological appearances of non-neoplastic breast tissues on pCLE (left panel) and corresponding histology images (right panel). We observe that there exists latent correspondence in view of visual similarity between pCLE and histology images.

tice. As an emerging technique, clinically the final diagnosis still needs the support from histology images.

Recently, there has been significant interest in designing classification-based or content-based retrieval model to determine the categorization of newly-sampled mosaics [5, 1, 3, 2, 22, 4]. SIFT [15] is widely used for visual representation of mosaics which is proved to outperform other features [5, 1]. Andre et.al [4] have introduced several semantic attributes to improve the retrieval accuracy and interpretability. However, we discovered several observations as follows:

Multi-view Representation Although SIFT-like features perform well in characterizing pCLE mosaics, they are designed to capture the local structures of images while the textures and contours cannot not be fully characterized. Several features including Texton [19] and HoG [8] are sound to gain additional information for visual representation. How to integrate multiple features of mosaics is one of core problem addressed in this paper.

Multi-modal Correspondence Current works on pCLE video retrieval and classification are mostly based on single modal dataset, i.e., they only use pCLE mosaics to design the strategies. However, the final diagnosis is often referenced to the histology images as a gold standard. The histology image seems to be able to provide extra information for better discrimination. However, the difference between mosaics and final histology is considerably large as shown in Figure 1. The latent relationship between mosaics and histology images exists but is difficult to uncover. Moreover, for real-time captured mosaics, the final histology image is not available. Therefore, another task in this paper is to design the latent correspondence between mosaics and histology images in offline learning which can be used in online schemes.

In this paper, we propose a *Multi-View Multi-Modal Embedding* (MVMME) framework to learn discriminative features of pCLE videos by exploiting both mosaics and histology images. For mosaic images, multiple features including SIFT, Texton and HoG are deployed as multi-view visual representations. For multimodal embedding, we propose a supervised scheme which generates a mapping from original features to a latent space by maximizing the semantic correlation between mosaics and histology images. The learned mapping function can transform multi-view mosaic representations into robust latent features.

The remainder of this paper is organized as follows: Section 2 briefly reviews the previous works on classification and retrieval approaches for pCLE videos as well as multi-view learning methods. In Section 4, the workflow of the proposed MVMME framework is introduced step by step. Empirical experiments on real datasets are conducted in Section 4. The conclusions and future works are presented in Section 5.

2. Related Works

2.1. Classification and Retrieval on pCLE

The research on pCLE video mostly relies on classification or retrieval tasks for determining the categorization of mosaics [5, 1, 3, 2, 22, 4]. In the work of Andre et. al[5], SIFT and nearest neighbor search are used for content-based image retrieval tasks with pCLE mosaics. In [1], Andre et. al proposed a densely-sampled SIFT approach for better representation as well as k-nearest neighbors classification with leave-one-patient-out (LOPO) cross-validation. [3] also follows the densely-sampled SIFT boosted by multi-scale detection. An effective metric is learned to calculate the similarity between bag-of-words features based on SIFT. In order to reduce the gap between visual content and diagnosis, Andre et. al introduced semantic attributes in [4] to describe the visual content that improve the performance on pCLE retrieval. Tafresh et. al [22] learned query-specific schemes to extract the region-of-interest (ROI) and relevant sub-sequence of video samples before pCLE video retrieval. Most of the previous works on pCLE video classification and retrieval tasks use only SIFT-like feature for visual representation. The information from other visual features and, more importantly, histology images has not been investigated.

2.2. Multi-view and Multiple Feature Learning

Multi-view learning is an active research topic in recent years. The multiple views can be the different viewpoints of an object in the camera, or the various descriptions of a given sample. Multi-view learning aims at matching different types of features or modalities to form a common shared subspace that maximises the cross-view correlation. A direct strategy is to concatenate the different kinds of features into a long vector. This often leads to the curse of dimensionality problem and thus it is not practical. Another typical approach to obtain a common space for two views is canonical correlation analysis (CCA) [11]. CCA is an unsupervised method that attempts to learn transforms by maximizing the cross correlation between two views. The extensions of CCA including MVCCA can process the multiple views problems [16, 17]. Supervised methods [18, 20, 12] learn the common space incorporating labels and categorization information. The class labels can be utilized to characterize the intra-class and inter-class discriminant.

The task for multiple feature fusion is to assign multiple features with weight in task-specific cases. For retrieval tasks, the early and late fusion are two representative approaches to integrate multiple features. For classification tasks, multiple kernel learning (MKL) strategies are widely used to fuse multiple kernels with support vector machines [9, 21]. MKL was also used for dimensionality reduction of the multi-view data based on graph embed-

ding [14]. Kloft et al. [13] extended the traditional L_1-norm MKL to arbitrary norms, and showed that the non-sparse MKL was superior to the state-of-the-art in combining different feature sets for biometrics recognition.

In this paper, we aim to design a supervised embedding strategy to represent mosaics with multiple weighted features and build a latent subspace between mosaics and histology images for discriminative representation.

3. Methodology

3.1. Problem Formulation and Denotations

The pCLE dataset is formed with a set of mosaic images $\{X_i\}, i = 1, \ldots, n$ where n is the number of mosaics. The mosaics are extracted from continuous videos which are labeled with two-level groundtruth. The coarse-level label L_i^c indicates whether the mosaic supports *non-neoplastic* or *neoplastic* status while the fine-level label L_i^f refers to fine-grained diagnosis and tissue information. Each mosaic is represented with multiple views of features that are denoted by $X_i^{(k)}, k = 1, \ldots, n_k$ where n_k is the number of visual features. A fraction of mosaics are matched with histology images $\{Y_i\}, i = 1, \ldots, n_h; n_h < n$. The task of the MVMME is to learn a mapping function from pCLE mosaics to latent space for discriminative representations exploiting multi-view multi-modal under semantic supervision.

3.2. Preprocessing and Visual Features

Since the mosaics are extracted from continuous pCLE videos, the size of region of interest (ROI) can vary with the capturing steps and the background as shown in the left side of Figure 3 will affect the performance of visual representation, especially the global features. We firstly crop the ROI from original mosaic as shown in the right side of Figure 3. Given the mosaics, we focus more on texture and contour of tissues. Therefore, visual features that can preserve the local texture information are deployed including SIFT, Texton and HoG.

The densely-sampled **SIFT** is claimed in previous works that can well capture the local features of pCLE mosaics. Therefore, we follow the similar settings in [1] to extract dense SIFT features for mosaics as well as histology images.

Texton feature [19] is based on a dense description of local texture features. Compared with SIFT that focuses on interesting points, Texton is extracted with a set of manually-designed filtering banks which allows the integration of task-specific priors.

HoG feature [8] has been widely used in computer vision tasks to characterize the local gradient information. Compared with SIFT and Texton, HoG is uniformly extracted with sliding window that reveals more local details. However, HoG cannot well promise the invariant of scale and rotation transformation.

Features mentioned above are designed to reflect the texture information in specific tasks. In this paper, we will address the issue of how to make fully use of these features to generate discriminative representation of mosaics.

3.3. Multi-view Multi-modal Embedding

To leverage the semantic labels for MVMME, we construct the pairwise semantic similarity based on two-level label vectors. More specifically, the similarity between the ith entity and the jth entity is defined as follows:

$$S_c(i,j) = \begin{cases} 0 & \text{if } (L_i^c)^T L_j^c = 0 \\ 1 + (L_i^f)^T L_j^f & \text{if } (L_i^c)^T L_j^c = 1 \end{cases} \quad (1)$$

According to the definition of S_c, when two samples belong to different classes in coarse-level, i.e. *non-neoplastic* or *neoplastic*, the similarity is zero. However, the diagnosis and tissue information is taken into consideration when samples belong to the same coarse-class.

Since the labels can provide semantic description for mosaics, the task for MVMME is to reconstruct the semantic similarity S_c in the latent subspace. We firstly focus on learning the mapping functions based on the mosaics with histology images in a single-view scheme where only single feature from mosaics and histology images are deployed. The objective function of the proposed method is to minimize the reconstruction error as follows:

$$\min_{f(\cdot),g(\cdot)} \sum_{i,j} (\frac{1}{c} f(X_i)^T g(Y_j) - S_c(i,j))^2 \quad (2)$$

where $f(\cdot)$ and $g(\cdot)$ are mapping functions to learn from original feature space to the latent subspace. c is the dimension of the latent subspace. Although many different kinds of functions can be used to define $f(\cdot)$ and $g(\cdot)$, we adopt the commonly used linear function form where $f(x) = W_x^T x$ and $g(y) = W_y^T y$. Therefore, the problem in Eq.(2) can be rewritten into the matrix form:

$$\min_{W_x, W_y} \|(XW_x)(YW_y)^T - cS_c\|_F^2 \quad (3)$$

The solution to problem in Eq.(3)can be obtained by adding orthogonality constraints and the transform matrix W_x can be learned. In order to deploy multiple visual features of mosaics, the objective function is extended as follows:

$$\min_{W_x^{(k)}, W_y} \|(\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)})(YW_y)^T - cS_c\|_F^2 \quad (4)$$

where $X^{(k)}$ is the kth feature of mosaics and $W_x^{(k)}$ is the corresponding transform function. In order to make the bits

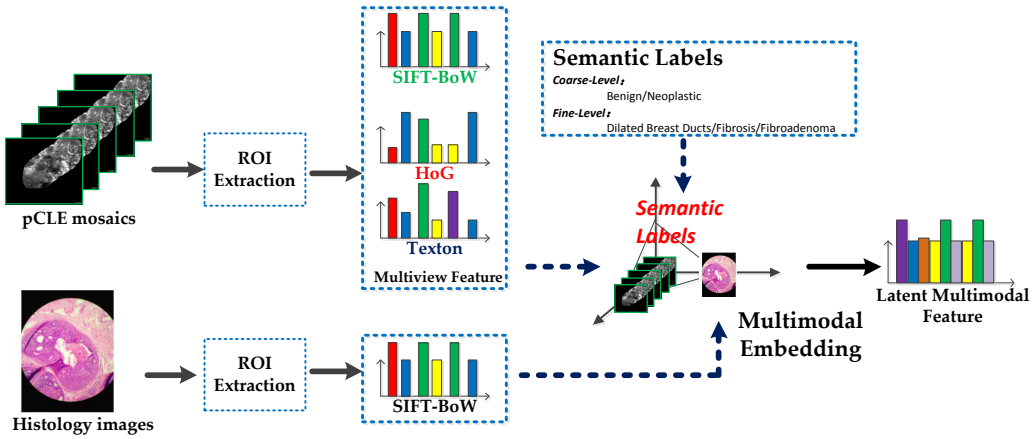
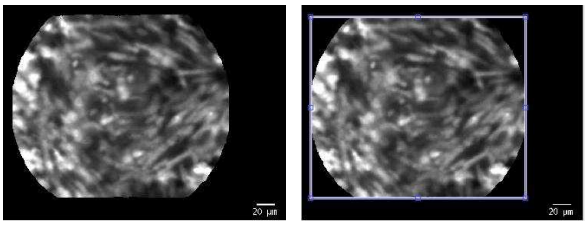


Figure 2. Workflow of the proposed framework: The regions of interest (ROIs) are firstly extracted from original mosaics and histology images. For mosaics, SIFT-BoW, HoG and Texton features are generated as visual representation while the histology images are represented with simple SIFT-BoW. A supervised embedding strategy maps the visual information from mosaics and histology images into a latent space exploiting two-level semantic labels. The learned feature can be further used a specific tasks.



(a) Original Frame (a) ROI Frame

Figure 3. Example of ROI detector

between different functions uncorrelated and preserve the local smoothness, we impose the manifold regularization and orthogonality constraints to Eq.(4). The final objective function is formulated as follows:

$$\begin{aligned}\min_{W_x^{(k)},W_y} &\|(\sum_{k=1}^{n_k} X^{(k)}W_x^{(k)})(YW_y)^T - cS_c\|_F^2 \\ &+ \sum_{i,j} Z_y(i,j)\|y_iW_y - y_jW_y\|^2 \\ &+ \sum_{i,j}\sum_{k=1}^{n_k} Z_x^{(k)}(i,j)\|\sum_{k=1}^{n_k} x_i^{(k)}W_x^{(k)} - x_j^{(k)}W_x^{(k)}\|^2 \\ s.t. &W_y^TY^TYW_y = nI_c \\ &(\sum_{k=1}^{n_k} X^{(k)}W_x^{(k)})^T(\sum_{k=1}^{n_k} X^{(k)}W_x^{(k)}) = nI_c\end{aligned} \tag{5}$$

where $Z_x^{(k)}(i,j) = e^{-\|x_i^{(k)}-x_j^{(k)}\|^2/\sigma}$ and $Z_y(i,j) = e^{-\|y_i-y_j\|^2/\sigma}$ are similarity matrix based on original features. With simple matrix transformation and substituting the Eq.(5), the above equation can be rewritten as follows:

$$\begin{aligned}\min_{W_x^{(k)},W_y} &\|(\sum_{k=1}^{n_k} X^{(k)}W_x^{(k)})(YW_y)^T - cS_c\|_F^2 \\ &+ tr(YW_yL_y(YW_y)^T) \\ &+ tr(\sum_{k=1}^{n_k} X^{(k)}W_x^{(k)}\sum_{k=1}^{n_k} L_x^{(k)}(\sum_{k=1}^{n_k} X^{(k)}W_x^{(k)})^T) \\ s.t. &W_y^TY^TYW_y = nI_c \\ &(\sum_{k=1}^{n_k} X^{(k)}W_x^{(k)})^T(\sum_{k=1}^{n_k} X^{(k)}W_x^{(k)}) = nI_c\end{aligned} \tag{6}$$

where $L_x^{(k)} = D_x^{(k)} - Z_x^{(k)}$ is the graph Laplacian of kth view of mosaics and $D_x^{(k)}$ is a diagonal matrix whose elements are the sum of each row in $Z_x^{(k)}$. Similarly, $L_y = D_y - Z_y$ is the graph Laplacian of histology images.

3.4. Optimization

In order to obtain the solution to Eq.(6), we denote the sum of manifold regularization terms as $L_R = tr(YW_yL_y(YW_y)^T) + \sum_{k=1}^{n_k} tr(X^{(k)}W_x^{(k)}L_x^{(k)}(X^{(k)}W_x^{(k)})^T)$ and then ex-

pand the objective function in Eq.(6) as follows:

$$\begin{aligned}
&\|(\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)})(YW_y)^T - cS_c\|_F^2 + L_R \\
&= tr[(\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)})(YW_y)^T (YW_y)(\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)})^T] \\
&- 2c \cdot tr[(\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)}) S_c (YW_y)^T] + tr(c^2 S_c^T S_c) + L_R \\
&= -2c \cdot tr[(\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)}) S_c (YW_y)^T] + L_R + const
\end{aligned}$$

where $tr(\cdot)$ denotes the trace of matrix. Therefore, the problem above can be converted into:

$$\begin{aligned}
\min_{W_x^{(k)}, W_y} \quad & - 2c \cdot tr[(\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)}) S_c (YW_y)^T] + \\
& + tr(YW_y L_y (YW_y)^T) \\
& + tr(\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)} \sum_{k=1}^{n_k} L_x^{(k)} (\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)})^T) \quad (7) \\
s.t. & W_y^T Y^T Y W_y = nI_c \\
& (\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)})^T (\sum_{k=1}^{n_k} X^{(k)} W_x^{(k)}) = nI_c
\end{aligned}$$

Then, let $\tilde{X} = [X^{(1)}, \ldots, X^{(n_k)}, Y]$ and :

$$\tilde{S} = \begin{bmatrix} \sum_{k=1}^{n_k} L_x^{(k)} & -cS_c \\ -cS_c & L_y \end{bmatrix}$$

$$\tilde{W} = \begin{bmatrix} W_x^{(1)}, \ldots, W_x^{(k)} & \mathbf{0} \\ \mathbf{0} & W_Y \end{bmatrix}$$

The final problem is formulated as follows:

$$\begin{aligned}
\min_{\tilde{W}} \quad & tr(\tilde{W}^T \tilde{X}^T \tilde{S} \tilde{X} \tilde{W}) \\
s.t. \quad & \tilde{W}^T \tilde{X}^T \tilde{X} \tilde{W} = 2nI_{2c}
\end{aligned} \quad (8)$$

It can be proved that the problem in Eq. (8) is equivalent to a generalized eigenvalue problem. The optimal solution of $\tilde{W}$ is the eigenvectors corresponding to the $2c$ smallest eigenvalues of $(\tilde{X}^T \tilde{S} \tilde{X})\tilde{W} = \lambda(\tilde{X}^T \tilde{X})\tilde{W}$. The projection matrix $W_x^{(k)}$ and W_y can also be obtained.

4. Experiment

4.1. Dataset and Measurement

The dataset is collected by a pre-clinical pCLE system (Cellvizio, Mauna Kea Technologies, Paris, France) as described in [7]. Breast tissue samples are obtained from 50 patients that are diagnosed with two classes at coarse-level including *non-neoplastic* and *neoplastic*. The fine-level labels are defined based on tissue and diagnosis information. The tissue information contains *adipose tissue*, *elastic fibres*, *collagen fibres* and *breast lobule*. The diagnosis result supports the existence of specific lesion including *ductal carcinoma in situ (DCIS)*, *invaisve ducal carcinoma (IDC)*, *invasive lobular carcinoma (ILC)*, *metaplastic carcinoma with spindle-cell morphology*, *invasive tubular carcinoma (ITC)* and *Invasice cancer infiltrating fat*. After completion of pCLE imaging, each sample underwent routine histopathology processing to generate the histology images. In order to evaluate the performance of MVMME, we conduct the classification based on coarse-level labels. The features from MVMME or other baseline approaches are fed into a support vector machine (SVM) classifier to determine the class.

Since the evaluation is exactly a two-class classification task, we use $Specificity$,$Sensitivity$ and $Accuracy$ and to measure the performance that can be calculated as follows:

$$\begin{aligned}
&Sensitivity = TP/(TP + FN) \\
&Specificity = TN/(TN + FP) \\
&Accuracy = (TN + TP)/(TN + FP + TP + FN)
\end{aligned}$$

4.2. Experimental Settings and Baselines

For visual representation, SIFT interesting points are extracted from mosaics and histology images respectively. We generate 500D BoW SIFT feature for mosaics and 200D BoW SIFT feature for histology images. All 46 blanks of Texton feature are extracted for each mosaics and a 1000D Bag-of-Words feature is obtained based on Texton. HoG features are only deployed in KF classification since the size of ROI is closed over all samples where the dimension of HoG is 97650D and then reduced to 128D by PCA [10]. The dimension of the latent space learned by MVMME is set to 10. Several baselines are implemented in this paper for comparison.

- **Single-view Raw Feature:** SIFT, Texton and HoG are individually used as visual representation and directly fed into the classifier. We denote these schemes with *SIFT-only*, *Texton-only* and *HoG-only*.

- **Single-view Multi-modal CCA**: The mosaic is represented by single feature that are embedded with histology images by unsupervised CCA [11]. We denote these schemes with *CCA-SIFT*, *CCA-Texton* and *CCA-HoG*.

- **Multi-view CCA**: The multi-view multi-modal embedding strategy implemented by Multi-view CCA that takes multiple features from mosaics and histology images equivalently. The result is denoted by *MVCCA*.

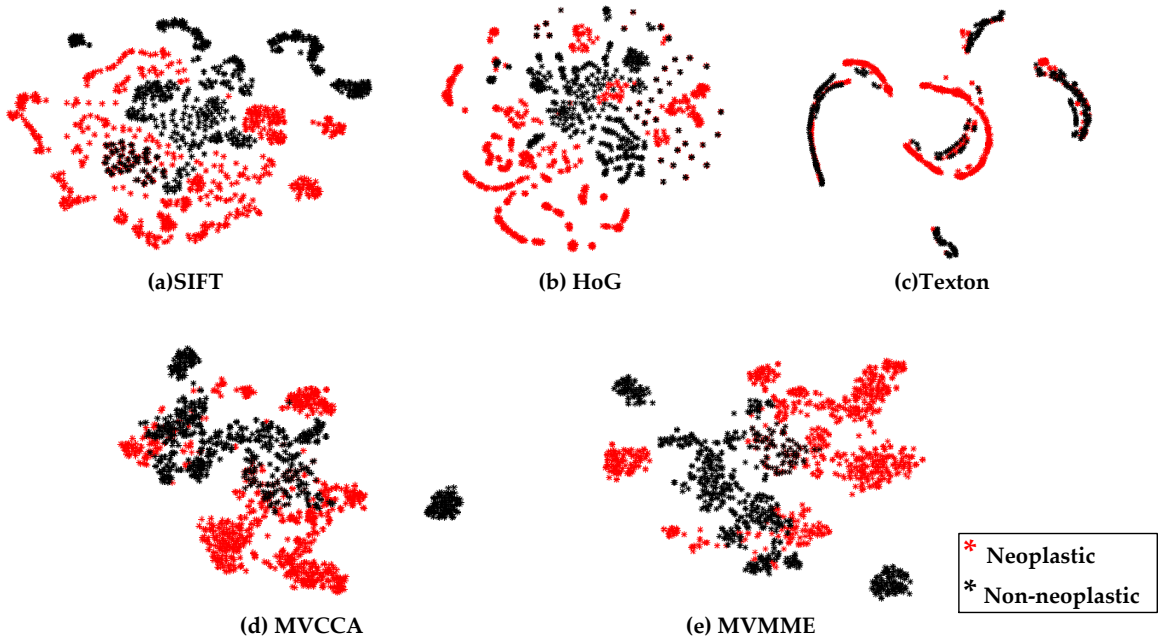


Figure 4. Visualization result by t-SNE: The red dots refer to the neoplastic samples while the black dots refer to non-neoplastic samples. Figures in the first line present the samples based on simple features including (a) SIFT, (b) HoG and (c) Texton. The embedded features by unsupervised MVCCA and the proposed supervised MVMME are listed in the second line.

- **Single-view MVMME**: Only one view of the mosaics is embedded with histology images via supervised mapping strategy proposed in this paper. The result is denoted by *MVMME-SIFT*, *MVMME-Texton* and *MVMME-HoG*. The proposed method that uses all features is denoted by *MVMME-All*.

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. We use LIBSVM package [6] for SVM classifier.

4.3. Visualization Results

We firstly provide the visualization of the raw feature and the embedded feature via t-SNE [23] that maps the original feature into a 2D space.

Figure 4 illustrates the distribution of mosaics with different type of visual representation. We can observe that there are overlaps between different classes when the mosaics are represented by original features. Therefore, the raw features are not discriminative. Although MVCCA can largely improve the performance as shown in Figure 4(d), the problem of overlapping samples is still not fully addressed. Figure 4(e) shows that the proposed MVMME makes the mosaics distributed in dense clusters as well as large margin between different classes. Therefore, the embedded feature are likely to provide better performance in separating the non-neoplastic and neoplastic cases.

4.4. Numerical Results

Table 1 presents the classification performance of multiple baseline approaches and the proposed MVMME. The following observations can be derived::

- Different features for mosaics gain different performance. Texton-based schemes have higher specificity while HoG and SIFT contribute to better sensitivity.
- Compared with single feature schemes, the multimodal embedding strategy that discover the latent representation between mosaics and histology images can largely improve the classification performance.
- The unsupervised MVCCA cannot fully make use of multiple views since it treats the features of mosaics and histology images equivalently. The semantic information is also not investigated to weight the features.
- The sensitivity of MVMME is 0.960 and final accuracy is 0.966 which achieves the best performance in comparison with baselines. The specificity is 0.972 which is ranked at 2nd place and close to the highest value obtained by CCA-SIFT.

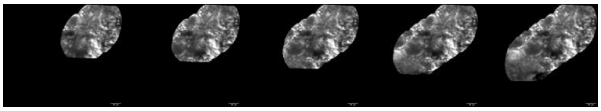
4.5. Key Frames v.s. Growing Frames

The experiments conducted above are based on the mosaics that are captured with increasing size of content as

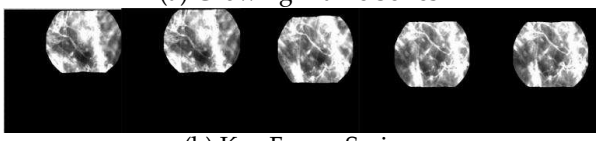
Table 1. Classification Accuracy

Method	$Sensitivity$	$Specificity$	$Accuracy$
Texton-Only	0.880	0.957	0.916
SIFT-Only	0.933	0.867	0.902
HoG-Only	0.876	0.833	0.853
CCA-Texton	0.889	0.940	0.913
CCA-SIFT	0.919	**0.989**	0.952
CCA-HoG	0.856	0.902	0.880
MVCCA	0.903	0.912	0.908
MVMME-Texton	0.937	0.943	0.938
MVMME-SIFT	0.923	0.972	0.956
MVMME-HoG	0.938	0.831	0.886
MVMME-All	**0.960**	0.972	**0.966**

shown in the upper side in Figure 5. In some cases, we only have the mosaics observed within the capturing scope which is shown in the lower side in Figure 5 denoted by KF. The key hole scheme moves the visible scope with approximately fixed size. Since the information is limited in KF, the classification is relatively challenging. We simply report the classification accuracy by single features (SIFT, HoG and Texton), MVCCA and MVMME in Table 2.



(a) Growing Frame Series



(b) Key Frame Series

Figure 5. Growing Frames (GF) contains the full procedure of capture. Key Frames (KF) only shows the scope of key holes.

Table 2. Classification Accuracy of KF and GF.

Method	GF	KF
SIFT	0.902	0.452
Texton	0.916	0.444
HoG	0.853	0.666
MVCCA	0.908	0.466
MVMME	0.966	0.652

According to Table 2, several results can be observed:

- The accuracy of GF classification is much higher than KF's as expected. Frames from non-neoplastic and neoplastic may share the same KF in different capturing progress which introduce ambiguity for classification. In contrast, the GF scheme provides complete view of tissues and the discriminative part can be preserved.
- Although the performance of KF is not promising, the integration of pCLE and histology by MVMME can largely boost the classification accuracy. Similarly, MVMME performs best in GF classification.

4.6. Discussion

According to the workflow of MVMME, parameters involved in the proposed method are limited except the dimension of the latent subspaces learned from mosaics and histology images. The dimension refers to c in Eq. (7) which is the number of smallest eigenvectors. Since the smallest eigenvalue of a matrix can be obtained with linear complexity, the scalability of the proposed method is promising. The classification accuracy with different dimensions of latent space is illustrated in Table 3.

Table 3. Classification Accuracy with different dimensions of the latent space.

Dimension	4	8	12
Accuracy	0.8228	0.8751	0.9750
Dimension	16	24	32
Accuracy	0.9841	0.9852	0.9886

According to the result in Table 3, with the dimension of the latent subspace increases, the classification accuracy is improved. However, when the dimension is larger than 16, MVMME cannot obtain significantly higher accuracy. Therefore, the proposed method can generate compact representation of mosaics that can largely reduce the storage space and the training time of classifiers.

5. Conclusion and Future Works

In this paper, we propose a *Multi-View Multi-Modal Embedding* (MVMME) framework to learn discriminative features of pCLE videos exploiting both mosaics and histology images. For mosaic images, multiple features including SIFT, Texton and HoG are deployed as multi-view visual representations. For multimodal embedding, we pro-

pose a supervised scheme which generates a mapping from original features to a latent space by maximizing the semantic correlation between mosaics and histology images. The learned mapping function can transform multi-view mosaic representations into robust latent features. The experiments on real dataset demonstrate that MVMME can outperform baseline approaches with single view or single modal features.

Since the dataset is not large and the number of labeled data is limited, the experiments on fine-level classification cannot be conducted. Moreover, the features used in this paper is relatively old-fashioned. In future works, the convolutional neural networks can be used to generate more powerful representations.

Acknowledgment

The authors thank Khushi Vyas and Tou Pin Chang for preparing the pCLE datasets used for this study. This research is partly supported by NSFC, China (No: 61572315, 61375048) and 973 Plan, China (No. 2015CB856004).

References

[1] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache. Endomicroscopic video retrieval using mosaicing and visualwords. In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, pages 1419–1422. IEEE, 2010. 2, 3

[2] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache. Retrieval evaluation and distance learning from perceived similarity between endomicroscopy videos. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pages 297–304. Springer, 2011. 2

[3] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache. A smart atlas for endomicroscopy using automated video retrieval. *Medical image analysis*, 15(4):460–476, 2011. 1, 2

[4] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache. Learning semantic and visual similarity for endomicroscopy video retrieval. *Medical Imaging, IEEE Transactions on*, 31(6):1276–1288, 2012. 2

[5] B. André, T. Vercauteren, A. Perchant, A. M. Buchner, M. B. Wallace, and N. Ayache. Endomicroscopic image retrieval and classification using invariant visual features. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 346–349. IEEE, 2009. 2

[6] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 6

[7] T. P. Chang, D. R. Leff, S. Shousha, D. J. Hadjiminas, R. Ramakrishnan, M. R. Hughes, G.-Z. Yang, and A. Darzi. 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):299–310, 2015. 5

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 2, 3

[9] Y. Gu, X. Qian, Q. Li, M. Wang, R. Hong, and Q. Tian. Image annotation by latent community detection and multikernel learning. *Image Processing, IEEE Transactions on*, 24(11):3450–3463, 2015. 2

[10] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933. 5

[11] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 2, 5

[12] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *Computer Vision–ECCV 2012*, pages 808–821. Springer, 2012. 2

[13] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011. 3

[14] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Dimensionality reduction for data in multiple feature representations. In *Advances in Neural Information Processing Systems*, pages 961–968, 2009. 3

[15] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. 2

[16] A. A. Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *Image Processing, IEEE Transactions on*, 11(3):293–305, 2002. 2

[17] J. Rupnik and J. Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, pages 1–4, 2010. 2

[18] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2160–2167. IEEE, 2012. 2

[19] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2, 3

[20] T. Sim, S. Zhang, J. Li, and Y. Chen. Simultaneous and orthogonal decomposition of data using multimodal discriminant analysis. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 452–459. IEEE, 2009. 2

[21] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006. 2

[22] M. K. Tafresh, N. Linard, B. André, N. Ayache, and T. Vercauteren. Semi-automated query construction for content-based endomicroscopy video retrieval. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pages 89–96. Springer, 2014. 2

[23] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008. 6