Coupled manifold learning for retrieval across modalities

Anees Kazi[1*] Sailesh Conjeti[1*] Amin Katouzian[2] Nassir Navab[1,3]

[1]Technische Universität München, Munich, Germany [2]IBM Research, Almaden, CA, USA
[3]Johns Hopkins University, Baltimore MD, USA

Abstract

Coupled Manifold Learning (CpML) is targeted at aligning data manifolds across two related modalities to facilitate similarity preserving cross-modal retrieval. Local and global topologies of the data cloud reflect intra-class variability and overall heterogeneity respectively making it critical to retain both for meaningful retrieval. Towards this we propose a learning paradigm which simultaneously aligns global topology while preserving local manifold structure. The global topologies are maintained by recovering underlying mapping functions in the joint manifold space by deploying partially corresponding instances. The inter-, and intra-modality affinity matrices are then computed to reinforce original data skeleton using perturbed minimum spanning tree (pMST), and maximizing the affinity among similar cross-modal instances, respectively. The performance of proposed algorithm is evaluated upon two benchmark multi-modal image-text datasets (Wikipedia and PascalVOC2012 - Sentence). We further show versatility and interdisciplinary application by extending it to cross-modal retrieval between multi-stain atherosclerosis histology medical image dataset. We exhaustively validate and compare CpML to other joint-manifold learning methods and demonstrate superior performance across datasets and tasks.

1. Introduction

Of late, multi-modal datasets such as text, images, videos *etc*. are growing widely. They are often descriptive of same concepts but are heterogeneous in nature, for *e.g.* web pages often contain illustrative images and associated textual information describing the image. Search and retrieval across these modalities is non-trivial as their metric spaces are often not comparable, termed as the *heterogeneity gap*. Keywords / Content-based retrieval often target similarity search within modality and do not seamlessly extend towards retrieval across them. Towards the same, in this work, we propose latent modality-invariant embeddings by Coupled Manifold Learning (CpML) that can be leveraged for similarity search for multi-modal datasets.

A reliable cross-modal image retrieval system is desirable as it carries immense potential in aiding decision-making by enabling access to all information across modalities that share semantic similarity. Specifically, within the medical imaging community, contrasting with single-modal image retrieval that has been an active research topic [19], the cross modal image retrieval has not yet been fully investigated for medical applications except for few works [8], [1], [5] that are also mainly adopted for health care management systems using text+image datasets. In the cross-modal retrieval task, the ultimate goal is to bridge the gap between feature spaces by mining their mutual correlations and unveiling similarities within a common latent space. To this end, several methods have been developed but the Canonical Correlation Analysis (CCA) [4] and its variants (ex. [11]) have been widely used for learning such a space by maximizing the correlation between the two feature spaces. Alternatively, learning coupled feature spaces (LCFS) [16] and Procrustes alignment [15] algorithms have been proposed, where the former focuses on selecting discriminative features while learning the subspace and the latter removes translational, rotational, and scaling components from one space so that the optimal alignment can be achieved. In general, majority of existing methods only preserve local geometries amongst features and ignore global geometries. In other words, they only ensure that similar instances in the original space become neighbors in the latent space but do not prevent dissimilar instances from being neighbors. Authors in [15] addressed this problem by projecting instances into latent space through recovered mapping functions upon partially corresponding instances and aligning the manifolds while preserving the global geometries.

The data across multiple modalities are inherently heterogeneous due to differences in their representation learning. Therefore, we need to preserve local structures that

*A. Kazi and S. Conjeti contributed equally as First authors.

carry information about population variability and at the same time preserve the global manifold geometries. This motivated us to develop the Coupled manifold learning (CpML) that respects both geometries in the latent space given partially corrsponding instances, which is accounted as the main contribution of this paper. This is achieved by: 1) incorporating perturbed Minimum Spanning Tree pMST [18] into CpML such that the original data skeleton is preserved and partially corresponding instances drive the alignment, and 2) introducing novel notion of proximity through inter- and intra-modality affinity matrices for maintaining local similarities within constructed graph neighborhood.

2. Methodology

In this section, we present the supporting mathematical formulation for CpML by considering two modalities $\mathcal{X}_1 = \left\{\mathbf{x}_1^i \in \mathbb{R}^{m_1}\right\}_{i=1}^{n_1}$ and $\mathcal{X}_2 = \left\{\mathbf{x}_2^i \in \mathbb{R}^{m_2}\right\}_{i=1}^{n_2}$ respectively. As a prerequisite, we collect $n_{\mathcal{L}}$ number of partially corresponding data (tuples) from both modalities constituting $\mathcal{L} \in \mathcal{X}_1 \times \mathcal{X}_2$. Depending on $\mathcal{L}$, we reconstitute $\mathcal{X}_1$ into two disjoint subsets: $\mathcal{X}_1^c$ and $\mathcal{X}_1^{wc}$ (with and without given tuples, respectively) and likewise partition $\mathcal{X}_2$. Typical cross-modal subspace learning algorithms leverage $\mathcal{L}$ *i.e.* $\mathcal{X}_1^c$ and $\mathcal{X}_2^c$ to mine correlations between the two modalities and effectively bridge the gap between the heterogeneous multi modal feature spaces by mapping them to a unified feature space. However, in scenarios of limited correspondence, just preserving neighborhood relationships amongst matching instances is of limited effect and can often over-fit thus limiting its generalization ability. CpML proposes to overcome this by using $\mathcal{X}_1^{wc}$ and $\mathcal{X}_2^{wc}$ together with $\mathcal{L}$ such that the whole global geometry of the two underlying manifolds couples and aligns in the joint feature space. For better understanding, we further divide this section into two parts as follows:

2.1. Graph Construction:

To facilitate cross-modal retrieval in $\mathcal{Z}$, similar data points across modalities should map closer and dissimilar points should be well separated. In the proposed method, preserving intra-modal similarity while projecting is achieved by discovering neighborhoods within modality (modeled as intra-modal perturbed minimum spanning trees (pMSTs)) and preserving them during CpML. Across-modality neighborhoods are inferred using $\mathcal{L}$ and modeled as links between the intra-modal pMSTs. During CpML, these 'links' aid in aligning the intra-modal pMSTs such that matching points across modalities are mapped close to one another in the unified space. Figure 1 illustrates the proposed CpML formulation for aligning two intra-modal pMSTs to learn the unified cross-modal space for retrieval.

`Step 1` ***Perturbed Minimum Spanning Tree (pMST)***: An ideal neighborhood graph should be representative of the underlying data manifold and its local structure. In contrast to naïve k nearest neighbor (k-NN) graph construction which is highly sensitive to the choice of k, the minimum spanning tree (MST) representation of the data manifold has desirable properties as it effectively represents the underlying skeleton of the manifold, does not introduce *gaps* between small random groupings of data points and theoretically guarantees connectedness of the graph. However, MST is too sparse and sensitive to noise. Alternatively, one can employ a fully-connected graph to represent the data distribution, but such an approach compromises on local neighborhoods and may introduce erroneous connections between traveling outside the underlying manifold.

To overcome the shortcomings of using a single MST, we employ an ensemble of MSTs generated from perturbed versions of the original data distribution, called perturbed Minimum Spanning Trees (pMST), which result in a '*reinforced*' skeleton of the underlying manifold which is more robust to noise and better representative of the local neighborhood structure. Given the original data $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, we generate perturbed copies $\mathcal{X}_p$ of $\mathcal{X}$ through a locally adaptive noise model *i.e.* $\mathcal{X}_p = \{\mathbf{x}_i^p \mid \mathbf{x}_i^p \in \mathcal{N}(\mathbf{x}_i, \sigma_i); \mathbf{x}_i \in \mathcal{X}\}$ where $\sigma_i = r_p \times d(\mathbf{x}_i, \mathbf{x}_i^k); r_p \in [0, 1]$ represents the locality of the noise model as it allows $\mathbf{x}_i$ to connect to different number of neighbors with each perturbed dataset. We generate $t_p > 1$ perturbed copies of the dataset using the earlier local noise model and fit an MST graph to each of them ($\texttt{MST}(\mathcal{X}_p)$). Edge e_{ij}^p between two points $\mathbf{x}_i^p$ and $\mathbf{x}_j^p$ takes a value of 1 if they are connected in $\texttt{MST}(\mathcal{X}_p)$ and 0 otherwise. pMST is an ensemble average of multiple MSTs generated for random perturbations of the data, where the edge weight $e_{ij} = \frac{1}{t_p}\sum_{p=1}^{t_p} e_{ij}^p$. For further use in defining the intra-modal proximity graph, we convert the pMST into a deterministic graph $\delta_{\mathcal{X}}$ with an edge if $e_{ij} > 0$. For CpML, we use the above definition of pMST to robustly identify only the neighborhood connections between amongst the points of $\mathcal{X}_1$ and $\mathcal{X}_2$, thus resulting in pMST models $\delta_{\mathcal{X}_1}$ and $\delta_{\mathcal{X}_2}$ for $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively. The weights along these edges is determined by the intra-modal distance metric described later.

`Step 2` ***Intra-modal Affinity***: For defining proper affinity and evaluating similarity/dissimilarity between data points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the manifold, we incorporate locally scaled l_2 norm into intra-modal distance metric $\mathcal{D}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 / (2\sigma_{\mathbf{x}_i}\sigma_{\mathbf{x}_j})$, where σ_i and σ_j are local scaling factors measured by $\sigma_i = \|\mathbf{x}_i - \mathbf{x}_K\|^2$ [17] such that $\mathbf{x}_K$ is the K^{th} neighbor of $\mathbf{x}_i$. This allows for self-tuning of point-to-point distances in local neighborhoods around the points $\mathbf{x}_i$ and $\mathbf{x}_j$. This formulation is used for calculating the intra-modal distance matrices $\mathcal{D}_{11}$ and $\mathcal{D}_{22}$ for $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively. In the particular case of cross-modal

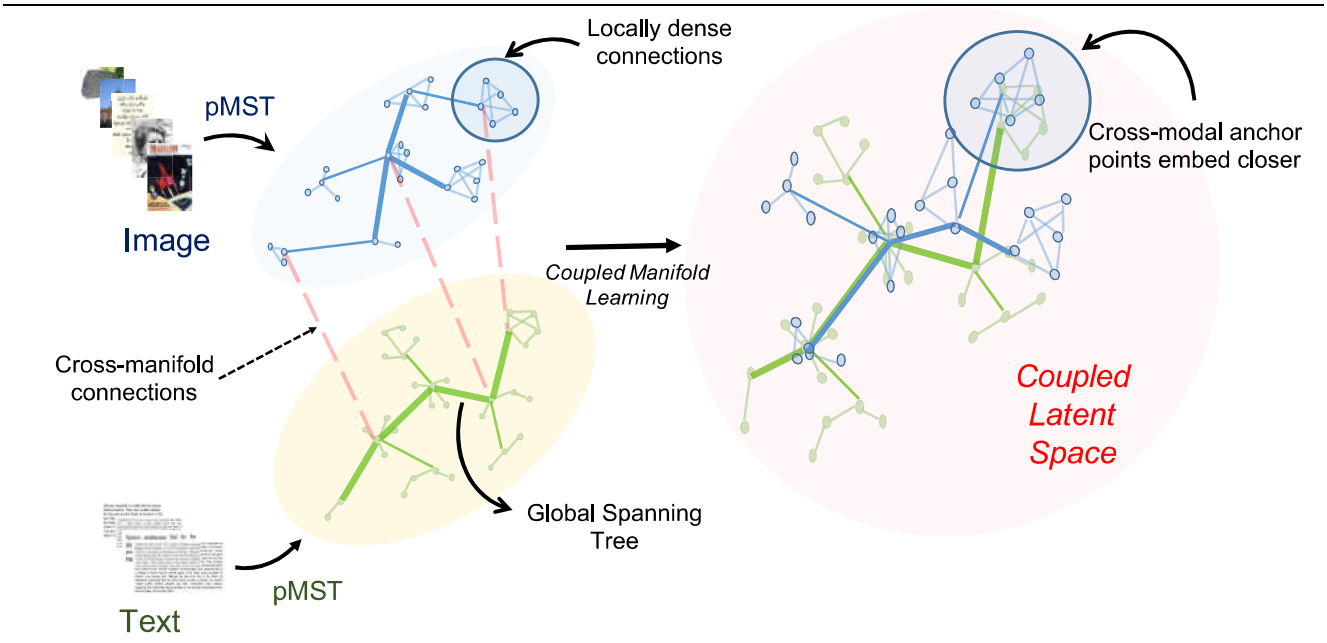

Figure 1: Schematic of proposed cross-modal search retrieval scheme using Coupled Manifold Learning (CpML). Given modalities (image and text) and limited co-occurring instances, we model the intra-modal proximity with pMST which creates locally dense connections with a global spanning tree representing the underlying data manifolds global topology. Next, we leverage correspondences through CpML, we learn to map to the coupled latent space that is makes the modalities metric-comparable.

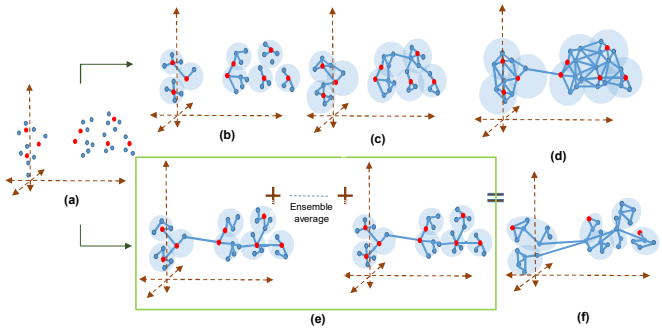

Figure 2: Schematic of comparatively illustrating potential graph construction paradigms on cloud of points (a). Graphs (b)-(d) are constructed using k-nearest neighbors. Graph (b) uses a small k and discovers local geometry well, however is not fully-connected in nature. Graph (c) uses moderate k, wherein local-geometry is traded-off gradually. Graph (d) uses a high value of k, however severely compromises on local geometry for global connectedness. Alternatively, using minimum spanning trees (MST) we obtain a fully-connected graph (e) with weak local support we propose to construct an ensemble of perturbed versions of MST (pMST) as shown in Graph (f) which has superior local support along with global connectedness representing the exoskeleton of the data manifold.

retrieval, the heterogeneous gap between the two features spaces warrants that we normalize the distance matrices $\mathcal{D}_{11}$ and $\mathcal{D}_{22}$ to make them comparable. Using normalized distances $\mathcal{D}_{11}$, $\mathcal{D}_{12}$ and neighborhood connections derived using the respective pMSTs, we define intra-modal affinity as $W_{11} = \exp(-\mathcal{D}_{11})$. $\delta_{\mathcal{X}_1}$ for $\mathcal{X}_1$ and likewise for $\mathcal{X}_2$.

`Step 3` ***Inter-modal Affinity***: To align the two manifolds (modalities) in the joint feature space, we need to compute affinities between instances across the modalities. We use the inferred intra-modal affinities W_{11}, W_{22} and given partial correspondences $\mathcal{L}$ to compute these affinities. We treat the corresponding instances across modalities as 'links' to align the modalities. For any pair of cross-modal points (say $\mathbf{x}_1^i$ and $\mathbf{x}_2^j$), the cross-modal affinity is computed as the maxima of affinities through all possible 'links' between the modalities, *i.e.* $W_{12}^{ij} = \max_{k \in [1, n_{\mathcal{L}}]} \sqrt{W_{11}^{ik} \times W_{22}^{kj}}$, where W_{11}^{ik} is the intra-modal affinity between $\mathbf{x}_1^i$ and $\mathbf{x}_1^k$, W_{22}^{kj} is the intra-modal affinity between $\mathbf{x}_2^j$ and $\mathbf{x}_2^k$ and $(\mathbf{x}_1^k, \mathbf{x}_2^k) \in \mathcal{L}$. We use the inferred affinity matrices (W_{11}, W_{22} and W_{12}) to construct the final composite distance matrix representing the joint geometry as: $\mathcal{D} = 1 - \begin{bmatrix} W_{11} & W_{12} \\ W_{12}^T & W_{22} \end{bmatrix}$.

2.2. Learning joint latent space:

The CpML casts the retrieval problem into learning a latent metric q-dimensional space $\mathcal{Z} = (\mathcal{Z}_{\mathcal{X}_1} \cup \mathcal{Z}_{\mathcal{X}_2} | \mathcal{X}_1 \rightarrow \mathcal{Z}_{\mathcal{X}_1}; \mathcal{X}_2 \rightarrow \mathcal{Z}_{\mathcal{X}_2}) \in R^q$ wherein $\mathcal{X}_1$ and $\mathcal{X}_2$ become comparable ($q \leq \min(m_1, m_2)$). CpML for retrieval between $\mathcal{X}_1$ and $\mathcal{X}_2$ translates to learning projection matrices $\alpha \in \Re^{m_1 \times q}$ and $\beta \in \Re^{m_2 \times q}$ defined for $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively. The learnt projection matrices transform $\mathcal{X}_1$ and $\mathcal{X}_2$ into the q-dimensional unified subspace $\mathcal{Z} = \left(\mathcal{Z}_{\mathcal{X}_1} \cup \mathcal{Z}_{\mathcal{X}_2} | \mathcal{Z}_{\mathcal{X}_1} = \mathcal{X}_1^T \alpha; \mathcal{Z}_{\mathcal{X}_2} = \mathcal{X}_2^T \beta\right)$. Learning linear projection matrices to align heterogeneous subspaces is preferred due to their ease of generalization to new unseen data samples and computational efficiency owing to direct mapping between feature space and the joint subspace.

The overall geometry comprising of both intra and inter-modal global geometries of the aligning manifolds can be modeled as a $(n_1 + n_2) \times (n_1 + n_2)$ joint distance matrix $\mathcal{D}$ representing the pairwise dissimilarity between any two instances in $\{\mathcal{X}_1, \mathcal{X}_2\}$. We use the definition of τ operator from [13] to uniquely characterize the joint manifold geometry, *i.e.* $\tau(\mathcal{D}) = -HSH/2$, where $S_{ij} = D_{ij}^2$ and $H_{ij} = \mathcal{I}^{(n_1+n_2)\times(n_1+n_2)} - (1/(n_1+n_2))$ where $\mathcal{I}$ is an identity matrix (The construction of $\mathcal{D}$ is discussed later). Ideally, the learnt latent space should preserve this global geometry and can be modeled as:

$$(\mathcal{Z}_{\mathcal{X}_1}^*, \mathcal{Z}_{\mathcal{X}_2}^*) = \underset{\mathcal{Z}_{\mathcal{X}_1}, \mathcal{Z}_{\mathcal{X}_2}}{\arg\min} \left\| \tau(\mathcal{D}) - \underbrace{[\mathcal{Z}_{\mathcal{X}_1}, \mathcal{Z}_{\mathcal{X}_2}]^T [\mathcal{Z}_{\mathcal{X}_1}, \mathcal{Z}_{\mathcal{X}_2}]}_{\mathcal{Z}^T \mathcal{Z}} \right\| \tag{1}$$

The above optimization problem can be re-posed as learning optimal projection matrices α^* and β^*, such that:

$$(\alpha^*, \beta^*) = \underset{\alpha,\beta}{\arg\min} \left\| \tau(\mathcal{D}) - k\,[\mathcal{X}_1\alpha, \mathcal{X}_2\beta]^T [\mathcal{X}_1\alpha, \mathcal{X}_2\beta] \right\|_2^2 \tag{2}$$

where k is a rescale factor and (α^*, β^*) are the optimal projection matrices. This can be re-stated as:

$$(\alpha^*, \beta^*) = \underset{\alpha,\beta}{\arg\min} \left\| \tau(\mathcal{D}) - Z^T f f^T Z | \right\| \tag{3}$$

where $Z = \begin{bmatrix} \mathcal{X}_1 & 0^{n_1 \times d_2} \\ 0^{n_2 \times d_1} & \mathcal{X}_2 \end{bmatrix}$ and $f = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$. Intrested readers are referred to [15] for a detailed mathematical derivation. To ensure that the projection vectors are not correlated and share similar dynamic ranges, we impose an additional sphering constraint on the optimization that $\begin{bmatrix} \alpha^T \mathcal{X}_1 & \beta^T \mathcal{X}_2 \end{bmatrix} \begin{bmatrix} \mathcal{X}_1^T \alpha \\ \mathcal{X}_2^T \beta \end{bmatrix} = \mathcal{I}^{q \times q}$.The optimal solution to Eq. 3 encourages preservation of local intra-modal neighborhoods and matching corresponding instances across modalities upon projection to the new joint space.

Learning Latent Space $\mathcal{Z}$ and Out of Sample Extension: There exists a optimal solution to the optimization problem posed in Eq. 1 using the eigendecomposition of $\tau(\mathcal{D})$ as $\tau(\mathcal{D}) = U^t \text{diag}(\Lambda_1, \cdots, \Lambda_q) U$ where $U \in \mathbb{R}^{(n_1+n_2)\times q}$. The latent subspace $\mathcal{Z}$ is estimated as $\mathcal{Z} = \text{diag}\left(\Lambda_1^{1/2}, \cdots, \Lambda_q^{1/2}\right) U$ [13]. For an unseen data point $\mathbf{x}_t$, we adapt the formulation from [12], which computes locally adaptive tangent spaces for out of sample extension (OSE). Through OSE, we seek the corresponding point $\mathbf{z}_t$ in the joint latent space $\mathcal{Z}$ and leverage the local neighborhood $\mathcal{N}(\mathbf{x}_t)$ defined in the high-dimensional space to define a locally linear mapping function M such that $\mathbf{z}_t = M\mathbf{x}_t$. M is decomposed into two piecewise matrices A and V ($M = AV$). V is inferred as the eigenvectors corresponding to the top q non-zero eigenvalues generated through Principal Components Analysis (PCA) on $\mathcal{N}(\mathbf{x}_t) \cup \mathbf{x}_t$. A is the similarity transformation matrix (translation, scaling and rotation) that is learnt through local Procrustes alignment.

Cross-modal Retrieval in Joint Space:Through CpML, we make the projected spaces $\mathcal{Z}_{\mathcal{X}_1}$ and $\mathcal{Z}_{\mathcal{X}_2}$ metric-comparable. Therefore, without loss of generality, the task of cross-modal retrieval for a query (say, $\mathbf{x}_q$ of modality $M1$) will be casted as projecting it appropriately onto the joint space ($\mathbf{z}_q = \text{OSE}(\mathbf{x}_q)$) and fetching the closest projected points from target modality ($\mathcal{Z}_{\mathcal{X}_2}$).

Extension to Feature Level Alignment: So far, the CpML was elaborated as it searches for and establishes non-linear mapping of original feature spaces and joint embedding space, which we refer to it as "instance-level" version (CpML-I). It can be seamlessly generalized to the case of linear embedding by replacing $\mathcal{Z}_{\mathcal{X}_1}$ and $\mathcal{Z}_{\mathcal{X}_2}$ in Eq. 1 with $\alpha^t \mathcal{X}_1$ and $\beta^t \mathcal{X}_2$, respectively. The solution is given by the eigenvectors corresponding to the q maximum non-zero eigenvalues of $Z\tau(\mathcal{D}) V^T \gamma = \lambda V V^T \gamma$ where $V = \begin{pmatrix} \mathcal{X}_1 & 0^{n_1 \times d_2} \\ 0^{n_2 \times d_1} & \mathcal{X}_2 \end{pmatrix}$ where $\gamma = [\alpha, \beta]$ [15]. This linear feature-level variant of CpML is thereafter refered to as CpML-F.

3. Experiments and Results

3.1. Datasets

To validate the versatility of the proposed method, we perform comparative analyses on two benchmark cross-modal retrieval datasets (Wikipedia and PascalVOC - Sentence) and extended CpML into the domain of medical image retrieval by demonstrating on cross-modal retrieval between multi-stain atherosclerosis histology datasets. We briefly describe the respective datasets below:

- *Wikipedia*: This dataset set consists of 2,866 im-

Algorithm 1: Instance-level Coupled Manifold Learning (CpML)

Input:
Training Data $\mathcal{X}_1 = \left\{\mathbf{x}_1^i \in \mathbb{R}^{d_1}\right\}_{i=1}^{n_1}, \mathcal{X}_2 = \left\{\mathbf{x}_2^i \in \mathbb{R}^{d_2}\right\}_{i=1}^{n_2}$;
Correspondence Tuples
$\mathcal{L} = \left\{\langle \mathbf{x}_1^i, \mathbf{x}_2^i \rangle \mid \mathbf{x}_1^i \in \mathcal{X}_1, \mathbf{x}_2^i \in \mathcal{X}_2, \mathbf{x}_1^i \leftrightarrow \mathbf{x}_2^i\right\}_{i=1}^{n_{\mathcal{L}}}$;
Parameter Set $\theta = \{\texttt{pMST} : r_p, t_p;$ Subspace Dimension : $q\}$;
Output: Projection matrices α, β

Step 1: Learn pMST
- S1.1: $\delta_{\mathcal{X}_1} = \texttt{pMST}(\mathcal{X}_1, r_p, t_p)$
- S1.2: $\delta_{\mathcal{X}_2} = \texttt{pMST}(\mathcal{X}_2, r_p, t_p)$

Step 2: Within Modality Affinity
- S2.1: $\mathcal{D}_{11}^{ij} = \frac{-\left\|\mathbf{x}_1^i - \mathbf{x}_1^j\right\|^2}{\sigma_{\mathbf{x}_1^j} \sigma_{\mathbf{x}_1^i}}$
- S2.2: $\mathcal{D}_{22}^{ij} = \frac{-\left\|\mathbf{x}_2^i - \mathbf{x}_2^j\right\|^2}{\sigma_{\mathbf{x}_2^j} \sigma_{\mathbf{x}_2^i}}$
- S2.3: $\mathcal{D}_{11}^{ij} = \texttt{Normalize}(\mathcal{D}_{11}^{ij})$
- S2.4: $\mathcal{D}_{22}^{ij} = \texttt{Normalize}(\mathcal{D}_{22}^{ij})$
- S2.5: $W_{11}^{ij} = \exp(-\mathcal{D}_{11}^{ij}).\ \delta_{\mathcal{X}_1}^{ij}$
- S2.6: $W_{22}^{ij} = \exp(-\mathcal{D}_{22}^{ij}).\ \delta_{\mathcal{X}_2}^{ij}$

Step 3: Across Modality Affinity
- S3.1: $W_{12}^{ij} = \max_{k \in [1, n_{\mathcal{L}}]} \sqrt{W_{11}^{ik} \times W_{22}^{kj}}$
- S3.2: $W_{12}^{ij} = (W_{12}^{ij} + W_{12}^{ji})/2$
- S3.3: $W_{21} = W_{12}'$

Step 4: Joint Graph Construction
- S4.1: $\mathcal{D} = 1 - \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$
- S4.2: $\mathcal{D} = \texttt{dijk}(\mathcal{D})$
- S4.3: $H = I - \frac{1}{(n_1+n_2)}$
- S4.4: Gram Matrix: $\tau(\mathcal{D}) = -H\mathcal{D}H/2$

Step 5: Estimating Projection Matrices
- S5.1: Define $Z = \begin{bmatrix} \mathcal{X}_1 & 0^{n_1 \times d_2} \\ 0^{n_2 \times d_1} & \mathcal{X}_2 \end{bmatrix}$
- S5.2: Let $\zeta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$
- S5.3: $\zeta^* = \arg\min_{\zeta} \|\tau(\mathcal{D}) - \tau(\mathcal{D}_{Z,\zeta})\|_2^2$ where $\mathcal{D}_{Z,\zeta} = Z^T \zeta \zeta^T Z$
- S5.4: Solution to S5.3 is given by eigenvectors corresponding to q largest eigenvalues of $Z\tau(\mathcal{D})Z^T\gamma = \lambda Z Z^T \gamma$
- S5.5: $\begin{cases} \alpha = \zeta^*(1 : m_1, 1 : q) \\ \beta = \zeta^*(m_1 + 1 : m_1 + m_2, 1 : q) \end{cases}$

return

age/text tuples curated into 10 semantic categories (split randomly into two disjoint training and testing datasets of 2,173 and 693 samples respectively). The images are represented using a SIFT codebook of 128 codewords and the text are modeled using a 10-topic Latent Dirchlet Allocation (LDA) model, trained unsupervised on a large corpus of images and texts respectively [9].

- *PASCAL - Sentence*: This dataset consists of 1000 images collected from 20 different categories of the PASCAL 2008 Challenge together with five descriptive sentences annotated using Amazon's Mechanical Turk. The images are described with a 2,790 dimensional feature vector which is a concatenated response of several object descriptors. Further, the corresponding textual annotations are described with a dictionary with 1,200 frequent and semantically relevant words to extract the feature vector for associated sentences [10].

- *Histology*: We followed an acquisition protocol in [6] and collected 253 HnE and MP pairs of cross-sections from 16 coronary arteries excised from 6 *post-mortem* human hearts, resulting in 16467 regions of interest (ROIs) with variable sizes (between 640μm $\times$ 640μm and 2560μm $\times$ 2560μm). The stains are performed on consecutive cross-sections ($< 5\mu$m apart) and rigidly registered manually. Eleven Modified AHA [14] labels were used for annotations of underlying tissues in accordance with interpretations from an expert cardiovascular histopathologist. It must be noted that CpML does not use labels during training and these annotations are used purely for validations. The ROIs were then fed into a pre-trained deeply learnt Convolutional Neural Network (CNN) trained for large-scale recognition tasks, to be purely used as a general-purpose *feature extractor*. Alternatively, one could train convolutional auto-encoder like architectures for representation learning. We used outputs arising from the penultimate fully connected layer of VGG-F [2] and AlexNet [7] deep CNN networks as 4096-dimensional features for HnE and MP images, respectively. Two different networks were chosen on purpose to maintain the heterogeneous gap between the raw feature spaces, which would subsequently be bridged through CpML and comparative methods. Further, to make the feature spaces discriminative, we reduced dimensionality using supervised locally linear projections, preserving 90% data variance [3].

3.2. Validation Scheme:

The performance of both CpML-I and CpML-F algorithms are evaluated against comparative methods as listed in Table 1. We randomly split the data into two disjoint subsets with a 80:20 ratio corresponding to the training and test datasets and repeated the splitting 10 times (For the histology-dataset, the split is generated artery-level). To evaluate sensitivity of CpML towards the need for cross-modal correspondences, we quantify the retrieval performance varying the degree of given correspondences for two settings of 20% (sparse) and 80% (dense) correspondences.

3.3. Results

The retrieval performance are measured using classification accuracy and for a particular query instance the class is predicted as the maximum a posteriori class evaluated from the top k nearest cross-modal neighbors. In Fig. 5, we demonstrate the classification accuracy through retrieval for the PASCAL and Wikipedia dataset for two settings text to image and image to text retrieval over the two datasets for two variations in the degree of correspondence (20% and 80%). The datasets had pre-determined test and train splits in a 20:80 ratio and these were maintained in this evaluation.

Table 1: Comparative methods and their configurational settings

Methods with abbreviations	Type	Graph	Hyperparameters
Cannonical Correlation Analysis (CCA) [4]	F	×	Cross-modal Correlation > 0.1
Manifold alignment preserving global geometry (MA-F and MA-I) [15]	F and I	FC	Eigen-value threshold $\epsilon > 10E-05$; k for OSE = 20
Learning coupled feature spaces (LCFS) [16]	F	×	Regularization parameters $\lambda_1 = 10E-01$,λ_2 = 10E-03 Number of iterations = 10
Procrustus Alignment (PA)	×	×	-
Cross-Modal Manifold Learning (CpML)	F and I	pMST	Number of perturbations $t_p = 20$ Locally adaptive noise model $r_p = 0.5; k = 5$ Eigen-value threshold $\epsilon > 10E-05$; k for OSE = 20

Table 2: Performance of comparative methods varying degree of correspondence given

		Methods	20%	40%	60%	80%	100%		Methods	20%	40%	60%	80%	100%
Histology Heterogenous	HnE → MP	CCA	25.08 ± 3.7	34.53 ± 4.3	40.75 ± 5.43	46.27 ± 3.1	50.12 ± 4.6	MP → HnE	CCA	30.57 ± 3.7	41.06 ± 4.52	47.63 ± 5.9	51.18 ± 3.8	56.24 ± 3.9
		MA-I	**13.69 ± 2.1**	**13.93 ± 1.8**	**14.25 ± 2.4**	**14.07 ± 1.8**	**14.07 ± 1.2**		MA-I	**19.39 ± 3.0**	**20.00 ± 5.2**	**19.36 ± 4.1**	**17.71 ± 3.7**	**16.67 ± 2.5**
		MA-F	32.94 ± 3.3	41.35 ± 3.4	45.52 ± 4.8	49.45 ± 3.1	48.84 ± 2.0		MA-F	46.06 ± 4.2	51.90 ± 4.4	57.02 ± 3.4	58.64 ± 4.8	59.15 ± 5.2
		LCFS	**51.73 ± 3.9**	61.18 ± 2.6	**65.89 ± 3.4**	68.12 ± 2.9	**69.92 ± 2.1**		LCFS	**66.12 ± 3.8**	**78.93 ± 3.2**	82.94 ± 1.7	83.93 ± 2.1	86.67 ± 1.4
		PA	37.63 ± 13.5	44.74 ± 17.6	48.41 ± 19.3	49.51 ± 21.7	51.71± 24.7		PA	44.30 ± 14.5	49.04 ± 19.4	52.42 ± 24.0	55.83 ± 23.9	55.59 ± 26.2
		CpML-I	47.80 ± 2.6	57.05 ± 2.8	64.94 ± 3.8	**68.46 ± 3.5**	**76.02± 2.2**		CpML-I	62.17 ± 5.1	**76.21 ± 5.3**	**85.57 ± 4.3**	**90.20 ± 1.1**	**95.23 ± 1.1**
		CpML-F	**62.37 ± 2.3**	**65.60 ± 3.0**	**65.26 ± 3.1**	**65.23 ± 1.7**	64.55 ± 1.8		CpML-F	**70.98 ± 4.6**	75.83 ± 3.7	75.43 ± 3.9	76.04 ± 1.8	75.13 ± 2.5

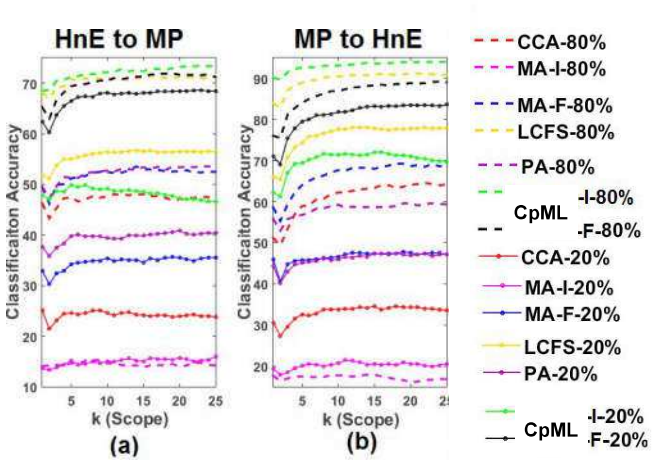

Figure 3: Performance *vs.* Scope (k retrieved cross-modal neighbors) curves for the proposed and comparative methods for 20% and 80% degrees of correspondence.

Fig. 3 depicts the overall performance, varying k through the accuracy-scope (k) curve for two settings of nearest neighbor retrieval (HnE → MP and MP → HnE) with 20% and 80% correspondences. Fig. 4 depicts the qualitative results of 3 cross-sections and corresponding retrived modality-couterpart images. The normal (N: left column), late fibroatheroma (FA: middle column), and pathological intimal thickening (PIT: right column) plaques have been successfuly retrived on the top 3 ranking results and only two are incorrectly fetched (red boxed) as the fourth neighbors. Such a retrieval tool will significantly improve histopathologist's ability to make reliable and fast decision.

Observations: From Fig. 3 and Table 2, we observe that the two proposed variants of CpML present a trend of consistently higher performance against comparative methods, substantiating the superiority of preserving joint global and local geometries. The performances of majority of methods are improved as degree of correspondences is increased from 20% to 80%. In case of CpML, this can be attributed to the better approximation of cross-modal affinity through given corresponding 'links' and hence making cross-modal data comparable in the latent space. Meanwhile, in majority of the cases, CpML-I shows improved performance over CpML-F due to the inherent non-linear flexibility of mapping data onto the embedding space.

The LCFS performance is closest to CpML as it discovers discriminative common latent space features, which make the embedding compact and effective. Additionally, despite considering global geometries while generating embedding, the MA-I underperformed, because, the Euclidean distance dissimilarity metric is not suitable for representing semantic similarity between instances.

4. Conclusions

We proposed CpML for effective cross-modal retrieval in which heterogeneous gap between cross-modal feature spaces is bridged by embedding instances into a metric-comparable latent space. In CpML, both local and global geometries are respected simultaneously using limited number of corresponding instances. The method has been benchmarked against state of the art methods and we demonstrate improved embedding, indirectly validated on standard text *vs.* image (and *vice versa*) cross-modal retrieval tasks. To the best of our knowledge, this is the first cross-modal medical image retrieval technique, demonstrating the versatility of CpML.

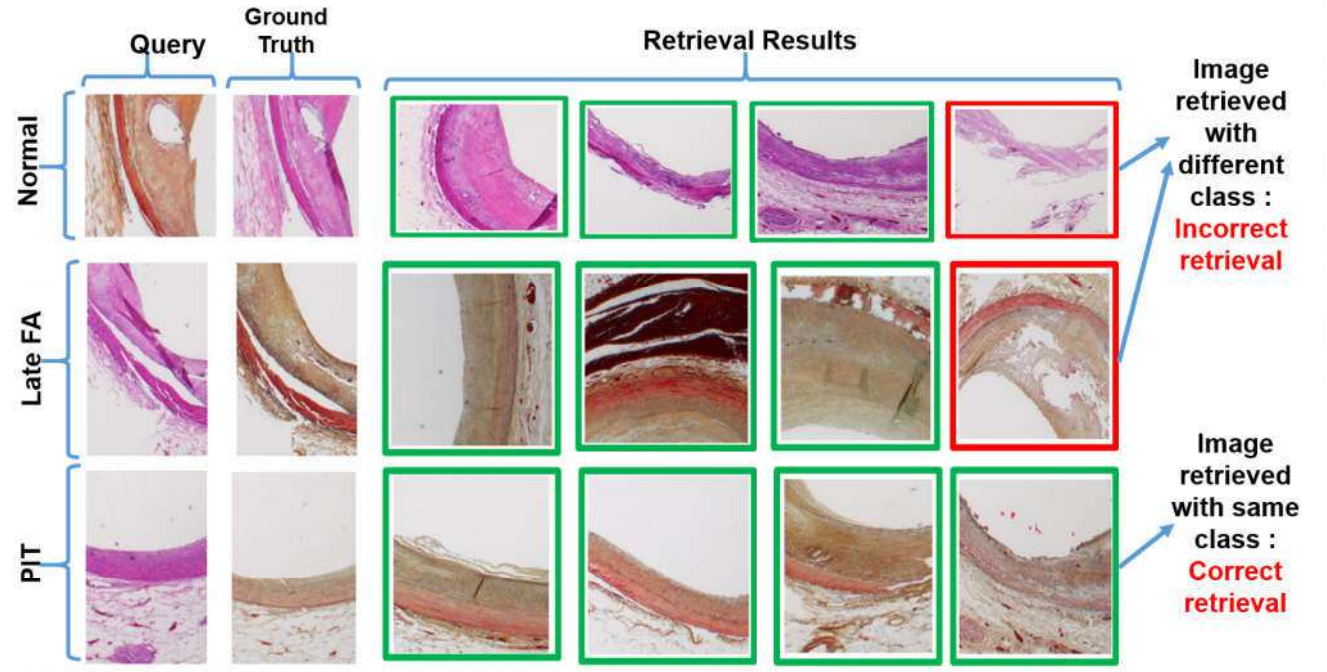

Figure 4: Qualitative results for atherosclerotic histology staging Query (Q) image along with 'ground truth' (GT) and top fetched cross-modal images (green box - similar annotation as Query and red box - dissimilar annotation).

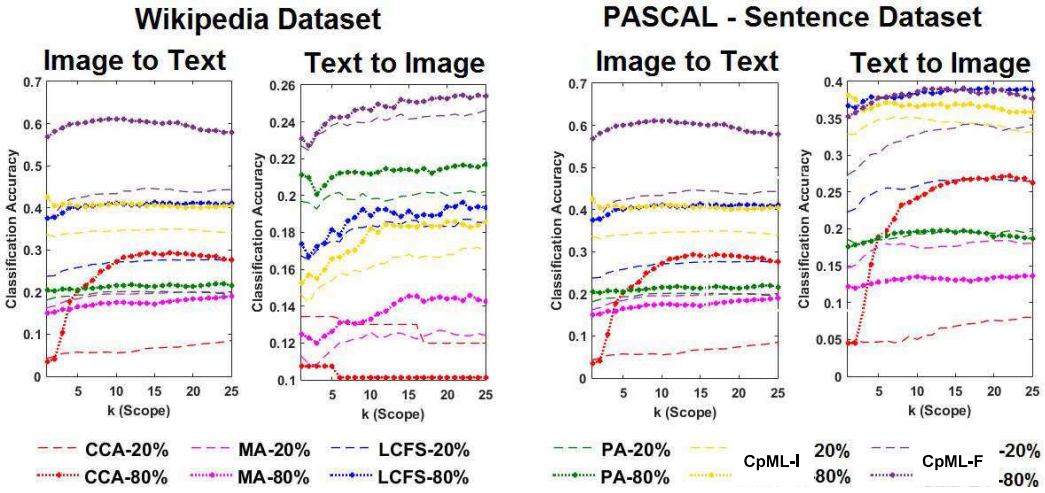

Figure 5: Performance of Cross-modal retrieval over public image-text datasets: Wikipedia and PASCAL sentence

.

References

[1] Y. Cao, S. Steffey, J. He, D. Xiao, C. Tao, P. Chen, and H. Müller. Medical image retrieval: a multimodal approach. *Cancer informatics*, 13(Suppl 3):125, 2014.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[3] X. He and P. Niyogi. Locality preserving projections. In *Advances in neural information processing systems*, pages 153–160, 2004.

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

[5] J. Kalpathy-Cramer, A. G. S. de Herrera, D. Demner-Fushman, S. Antani, S. Bedrick, and H. Müller. Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at imageclef 2004–2013. *Computerized Medical Imaging and Graphics*, 39:55–61, 2015.

[6] A. Katouzian, S. G. Carlier, and A. F. Laine. Methods in atherosclerotic plaque characterization using intravascular ultrasound images and backscattered signals. In *Atherosclerosis Disease Management*, pages 121–152. Springer, 2011.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] A. Kumar, J. Kim, W. Cai, S. Eberl, and D. Feng. A graph-based approach to the retrieval of dual-modality biomedical images using spatial relationships. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 390–393. IEEE, 2008.

[9] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.

[10] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.

[11] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.

[12] H. Strange and R. Zwiggelaar. A generalised solution to the out-of-sample extension problem in manifold learning. In *AAAI*, pages 293–296, 2011.

[13] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[14] R. Virmani. Lessons from sudden coronary death: a comprehensive morphological classification scheme for atherosclerotic lesions. *DIALOGUES IN CARDIOVASCULAR MEDICINE*, 10(3):189, 2005.

[15] C. Wang and S. Mahadevan. Manifold alignment preserving global geometry. In *IJCAI*, pages 1743–1749, 2013.

[16] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2095, 2013.

[17] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2005.

[18] R. S. Zemel and M. A. Carreira-Perpinán. Proximity graphs for clustering and manifold learning. In *Advances in neural information processing systems*, pages 225–232, 2005.

[19] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang. Towards large-scale histopathological image analysis: Hashing-based image retrieval. *IEEE Transactions on Medical Imaging*, 34(2):496–506, 2015.