Parallel Field Alignment for Cross Media Retrieval

Xiangbo Mao
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, China
mao.stingrey@gmail.com

Binbin Lin
The Biodesign Institute
Arizona State University
Tempe, US
binbin.lin@asu.edu

Deng Cai
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, China
dengcai@gmail.com

Xiaofei He
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, China
xiaofeihe@gmail.com

Jian Pei
School of Computing Science
Simon Fraser University
Burnaby, Canada
jpei@sfu.ca

ABSTRACT

Cross media retrieval systems have received increasing interest in recent years. Due to the semantic gap between low-level features and high-level semantic concepts of multimedia data, many researchers have explored joint-model techniques in cross media retrieval systems. Previous joint-model approaches usually focus on two traditional ways to design cross media retrieval systems: (a) fusing features from different media data; (b) learning different models for different media data and fusing their outputs. However, the process of fusing features or outputs will lose both low- and high-level abstraction information of media data. Hence, both ways do not really reveal the semantic correlations among the heterogeneous multimedia data. In this paper, we introduce a novel method for the cross media retrieval task, named Parallel Field Alignment Retrieval (PFAR), which integrates a manifold alignment framework from the perspective of vector fields. Instead of fusing original features or outputs, we consider the cross media retrieval as a manifold alignment problem using parallel fields. The proposed manifold alignment algorithm can effectively preserve the metric of data manifolds, model heterogeneous media data and project their relationship into intermediate latent semantic spaces during the process of manifold alignment. After the alignment, the semantic correlations are also determined. In this way, the cross media retrieval task can be resolved by the determined semantic correlations. Comprehensive experimental results have demonstrated the effectiveness of our approach.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis; H.3.3 [**Information Search and Retrieval**]: Retrieval Models

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502087.

General Terms

Algorithms

Keywords

Cross media, manifold alignment, parallel field, text corpus and images, multimedia

1. INTRODUCTION

In recent years, multimedia contents including text, image, audio and video on the web have been growing rapidly. With the role switch between social media (Twitter, Facebook) and traditional media (newspapers, etc.), more and more multimedia contents are published on the web by people. However, the explosion of multimedia contents has not been matched by an equivalent increase in the sophistication of multimedia content retrieval technology [7, 24, 25]. Nowadays, the dominate search engines for multimedia retrieval, such as Google and Bing, are still text-based. To effectively leverage the massive explosion of multimedia content, a large number of approaches have been proposed in the areas such as information retrieval, multimedia retrieval and computer vision [2, 3, 5, 11, 16]. One of the well known challenges in the area of multimedia retrieval is the so called semantic gap, *i.e.*, low-level features are not sufficient to characterize high-level semantics of media data [34, 35]. Aiming at this point, many previous works [27] have been proposed to simultaneously utilize multiple types of information such as the original multimedia contents, surrounding texts (or labels), and links to improve the multimedia retrieval performance. However, these works do not consider the semantic correlation among different media types. Generally, they can be viewed as uni-media retrieval, in which the query example and the retrieved results are of the same media type [7].

It is common knowledge that an important requirement for further progress in these areas is the development of sophisticated joint models for multiple media types. In which the most significant is the development of models that support inference with respect to content that is rich in multiple media types [24]. Specifically, these models should utilize the full structure of document which has a body of text accompanied with images or videos. For example, wikipedia page or newspaper article usually contains several paragraphs of

text and a number of related images for illustration. The performance of such models is referred as a cross media retrieval problem: the retrieval of all documents with the other media types in response to a media type query data. The task is crucial to many practical applications, such as finding images on the web that best illustrate a given text, finding the texts which are most related to a given image, or further, searching using a combination of different types of multimedia data and labelling media data automatically [24]. Figure 1 illustrates the following example, given an image of a horse, the cross media retrieval model should return all the contents related to the horse concept, *e.g.*, the sound of horse, the introduction text.

Figure 1: An example of cross media retrieval model.

To address the cross media retrieval problem, advances have been reported over the last decades [7, 26, 28]. These methods focus on two traditional ways to design cross media retrieval systems: (a) fusing features from different media data into a single vector [23, 33]; (b) learning different models for different media data and fusing their outputs [14, 32]. And most of these approaches require multiple-type queries, *e.g.*, queries composed of both image and text features. Hence, these methods are extensions of the classic uni-media retrieval systems.

The key challenge of cross media retrieval is to explore the semantic correlations among the heterogeneous media data. However, both of previous ways do not really reveal the semantic correlations. Semantic correlations can help us to better understand, organize and make use of the media data [34].

Recently, some developments bring new perspective to solve the cross media retrieval problem. Multimedia Correlation Space [34] and Correlation Semantic Space [24] introduce the idea of constructing a joint model to project original multimedia features to the semantic correlation space. And in the same period, manifold alignment methods [10, 29, 31] are proposed and shown that they are appropriate joint models for pair matching between heterogeneous data sources. However, most of existing manifold alignment methods use graph based regularizer *e.g.*, graph Laplacian, which focus on ensuring the first order smoothness of the mapping functions [21] in manifold alignment process. The first order smoothness of the mapping functions is not enough to reveal the underlying semantic correlations between heterogeneous types of multimedia data. In order to discover the latent semantic correlations, we would like to ensure the second order smoothness of the mapping functions in manifold alignment which preserve the geodesic distance of the manifolds. In some recent work, parallel fields [22] are found capable to keep the second order smoothness of the mapping functions [18].

Inspired by these developments in the cross media retrieval area, we propose a novel approach for cross media retrieval, called Parallel Field Alignment Retrieval (PFAR), which integrates a manifold alignment framework from the perspective of vector fields. Instead of fusing original features or outputs, we consider the cross media retrieval as a manifold alignment problem using parallel fields. The proposed manifold alignment algorithm can effectively preserve the metric of data manifolds, model heterogeneous media data and project their relationship into intermediate latent semantic spaces during the process of manifold alignment. After the alignment, the semantic correlations are also determined. In this way, the cross media retrieval task can be resolved by the determined semantic correlations. The empirical results from a real world data demonstrate the benefits of our approach over state-of-the-art cross media retrieval methods.

The rest of this paper is organized as follows. In Section 2, we introduce some basic background information about manifold alignment and parallel fields. In Section 3, we describe the proposed cross media retrieval algorithm (PFAR) in great detail. The experimental results of real world cross media data are presented in Section 4. Finally, Section 5 provides some concluding remarks.

2. BACKGROUND

As discussed in the previous section, our cross media retrieval approach involves manifold alignment and parallel fields techniques. In this section, we introduce some background on mainfold alignment and parallel fields.

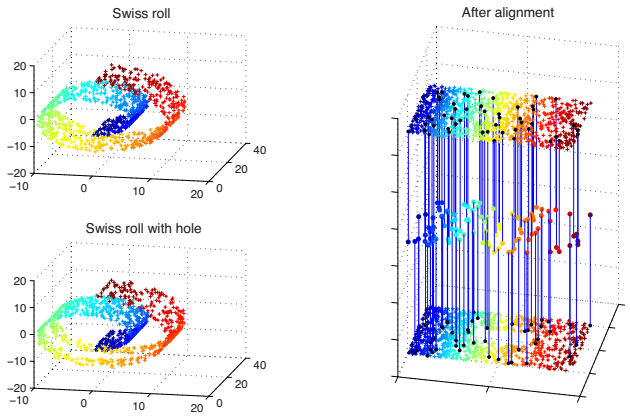


Figure 2: An example of manifold alignment. Two data manifold samples are shown in the left.The result of these two manifolds after alignment is shown in the right.

2.1 Manifold alignment

In many areas of machine learning and data mining, one is often confronted with very high dimensional data such as high definition videos and large vector-space documents. Learning problems involving these datasets are usually challenging. However, in many cases, there is a strong intuition that the high dimensional data may have a lower dimensional intrinsic representation. Manifold alignment is a class of machine learning algorithms which takes advantage of this assumption to produce projections between sets of data by aligning their underlying manifold representations [9, 30].

One of the pioneering work in manifold alignment is the paper of semi-supervised alignment [10]. Given certain labeled samples, semi-supervised alignment aims to find two maps which map two datasets to the new common space while satisfying the label constraint. Suppose $\mathbb{U}$ is the vertex space, there are l labeled vertices $\mathcal{V}_{label} = \{v_1, v_2, \ldots, v_l\}$, $\mathcal{V}_{label} \subset \mathbb{U}$, and $\mathcal{T}_{label} = \{t_1, t_2, \ldots, t_l\}$ ($t_i \in \mathbb{R}$), which is the vector of labeled target values. Similar to regression models, we would like to find a map defined on the vertices of the graph $f : \mathbb{U} \to \mathbb{R}$ which matches known target values for the labeled vertices. This can be solved by minimizing the following objective function:

$$E(f) = \sum_i \mu |f(v_i) - t_i| + f^T L f \tag{1}$$

Here $v_i \in \mathcal{V}_{label}$, $t_i \in \mathcal{T}_{label}$, and L is the graph Laplacian matrix. The relative weighting of these terms is given by the coefficient μ. The symmetric graph Laplacian matrix $L = L^T$ provides the information of the data manifold structure. Given two datasets with corresponding labels, we can learn manifolds for each datasets according to Equation 1, and align these two manifold with the intrinsic coordinates of labels. An illustration of manifold alignment is shown in Figure 2.

2.2 Parallel field regularization

Given a manifold $\mathcal{M}$, a vector field is a mapping from the manifold to tangent spaces on the manifold [22]. We can think of a vector field on the manifold $\mathcal{M}$ in the same way as we think of the vector field in Euclidean space. For each point on the manifold, we assign an arrow on it, with a given magnitude and direction, chosen to be tangent to the manifold $\mathcal{M}$. A smooth vector field means that tangent vectors vary smoothly on the manifold. An example of vector fields is shown in Figure 3.

One kind of the most important vector fields are parallel fields. The definition of parallel fields on the manifold is given as follows.

Definition 1. (Parallel Fields [22]) A vector field X on the manifold $\mathcal{M}$ is a parallel field if

$$\nabla X \equiv 0,$$

where ∇ is the covariant derivative on $\mathcal{M}$.

The parallel field is closely related to the linear function on the manifold. Let V be a parallel field on the manifold. If it is also a gradient field for function f, $V = \nabla f$, then f must be a linear function on the manifold.

Parallel fields essentially captures the second order smoothness of functions on the manifold [18]. Some recent theoretical results [15] shows that penalizing the second order

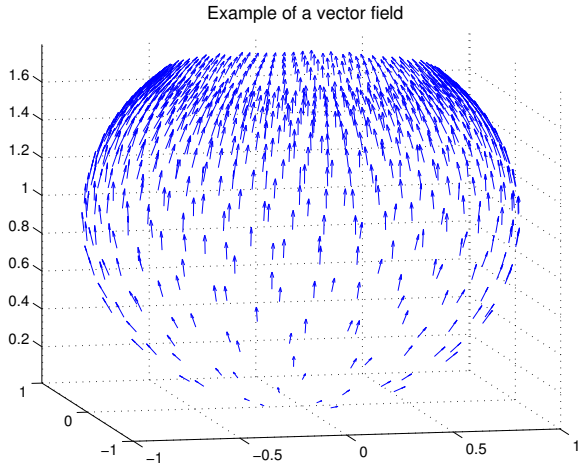


Figure 3: A vector field on punched sphere.

smoothness of functions helps achieve faster rates of convergence for semi-supervised regression problems. Moreover, parallel fields can also captures the metric structure of the manifold [17]. In other words, we can use parallel fields to preserve the distance on the manifold.

Next we briefly introduce Parallel Field Regularization (PFR). Let $\mathcal{M}$ be a d-dimensional manifold embedded in Euclidean space $\mathbb{R}^m$. Given l labeled data points $(x_i, y_i)_{i=1}^l$, $x_i \in \mathcal{M}$ and $y_i \in \mathbb{R}$, the goal of semi-supervised regression on the manifold is to learn a function $f : \mathcal{M} \to \mathbb{R}$. Specifically, PFR tries to learn the function f and it gradient field ∇f simultaneously via regularization. Formally, PFR aims to learn a function f and a vector field V by optimizing the following objective function:

$$E(f, V) = \frac{1}{l} \sum_{i=1}^{l} R_0(x_i, y_i, f) + \lambda_1 R_1(f, V) + \lambda_2 R_2(V), \tag{2}$$

where

$$R_1(f, V) = \int_{\mathcal{M}} \|\nabla f - V\|^2 \tag{3}$$

and

$$R_2(V) = \int_{\mathcal{M}} \|\nabla V\|_F^2. \tag{4}$$

The first term in Equation 2 is the loss function, and the second term enforces the vector field V to be close to the gradient field ∇f of f. ∇V in the third term measures the change of the vector field V. If ∇V vanishes, V must be a parallel field.

3. CROSS MEDIA RETRIEVAL

In this section, we present a novel approach to solve cross media retrieval problem.

3.1 The problem

Suppose we have a dataset $\mathcal{X} = \{X_1, \ldots, X_{|\mathcal{X}|}\}$ which contains documents of two different types of media data $\mathcal{A}$ and $\mathcal{B}$, *e.g.*, $\mathcal{A}$ is a collection of texts and $\mathcal{B}$ consists of images. Specifically, all X_i, $i \in (1, |\mathcal{X}|)$, in $\mathcal{X}$ can be quite diverse: from documents where a single text is complemented

by one or more images to documents containing multiple images and texts. For simplicity, we consider the case where each document in $\mathcal{X}$ consists of two sample components, one is from $\mathcal{A}$ and the other one is from $\mathcal{B}$, *i.e.* $X_i = (\mathcal{A}_i, \mathcal{B}_i)$. All data points of $\mathcal{A}$ and $\mathcal{B}$ are of vector forms in feature spaces $\mathcal{R}^{\mathcal{A}}$ and $\mathcal{R}^{\mathcal{B}}$, respectively. Under this circumstances, each document establishes a one-to-one mapping between component from the $\mathcal{A}$ and component from $\mathcal{B}$ media data spaces.

We consider cross media retrieval problem based on the doument dataset $\mathcal{X}$. The fundamental concept of cross media retrieval is rather straightforward. Given a query $Q_{\mathcal{A}} \in \mathcal{R}^{\mathcal{A}}$ (or $Q_{\mathcal{B}} \in \mathcal{R}^{\mathcal{B}}$), the goal of cross media retrieval is to return the closest matches in the $\mathcal{B}$ (or $\mathcal{A}$) space $\mathcal{R}^{\mathcal{B}}$ (or $\mathcal{R}^{\mathcal{A}}$). Whenever the $\mathcal{A}$ and $\mathcal{B}$ media data spaces have a natural correspondence, the original cross media retrieval problem can reduce to a classical retrieval problem: finding an invertible mapping function f between $\mathcal{A}$ and $\mathcal{B}$:

$$f : \mathcal{R}^{\mathcal{A}} \to \mathcal{R}^{\mathcal{B}}. \tag{5}$$

Hence, if a query $Q_{\mathcal{A}} \in \mathcal{R}^{\mathcal{A}}$ is given, we would be able to find the nearest neighbor of $f(Q_{\mathcal{A}})$ in $\mathcal{R}^{\mathcal{B}}$. Similarly, given a query $Q_{\mathcal{B}} \in \mathcal{R}^{\mathcal{B}}$, it would be easy for us to find the nearest neighbor of $f^{-1}(Q_{\mathcal{B}})$ in $\mathcal{R}^{\mathcal{A}}$ [24].

However, since $\mathcal{A}$ and $\mathcal{B}$ are different types of media data, they tend to adopt different feature representations. Therefore, it is hard to reveal semantic correlations between $\mathcal{R}^{\mathcal{A}}$ and $\mathcal{R}^{\mathcal{B}}$, which means the semantic gap still exists. For example, suppose $\mathcal{A}$ is a dataset consists of texts and $\mathcal{B}$ is a dataset consists of images. And we adopt TF/IDF and Histogram of Oriented Gradients (HOG) [4] to be the feature representations for texts and images, respectively. Thus, it is hard for us to directly explore the semantic correlations between text and image data spaces.

In order to abridge the semantic gap between $\mathcal{R}^{\mathcal{A}}$ and $\mathcal{R}^{\mathcal{B}}$, it is possible to map these two representations into two intermediate spaces which have a natural correspondence [24] and semantic correlations [36]. Consider following mappings:

$$f_{\mathcal{A}} : \mathcal{R}^{\mathcal{A}} \to \mathcal{I}^{\mathcal{A}}, \tag{6}$$

and

$$f_{\mathcal{B}} : \mathcal{R}^{\mathcal{B}} \to \mathcal{I}^{\mathcal{B}}. \tag{7}$$

Here, $f_{\mathcal{A}}$ and $f_{\mathcal{B}}$ map original $\mathcal{R}^{\mathcal{A}}$ and $\mathcal{R}^{\mathcal{B}}$ media data spaces to a pair of intermediate spaces $\mathcal{I}^{\mathcal{A}}$ and $\mathcal{I}^{\mathcal{B}}$, respectively. Further, we would like to ensure that there is also an invertible mapping between $\mathcal{I}^{\mathcal{A}}$ and $\mathcal{I}^{\mathcal{B}}$:

$$f_{\mathcal{I}} : \mathcal{I}^{\mathcal{A}} \to \mathcal{I}^{\mathcal{B}}. \tag{8}$$

Now, if a query $Q_{\mathcal{A}} \in \mathcal{R}^{\mathcal{A}}$ is given, the cross media retrieval task becomes finding the nearest neighbor of $f_{\mathcal{B}}^{-1} \circ f_{\mathcal{I}} \circ f_{\mathcal{A}}(Q_{\mathcal{A}})$ in $\mathcal{R}^{\mathcal{B}}$. Similarly, the goal becomes finding the nearest neighbor of $f_{\mathcal{A}}^{-1} \circ f_{\mathcal{I}}^{-1} \circ f_{\mathcal{B}}(Q_{\mathcal{B}})$ in $\mathcal{R}^{\mathcal{A}}$ if a query $Q_{\mathcal{B}} \in \mathcal{R}^{\mathcal{B}}$ is given [24]. Under this circumstances, the original problem of cross media retrieval is equivalent to learn the intermediate spaces $\mathcal{I}^{\mathcal{A}}$ and $\mathcal{I}^{\mathcal{B}}$.

In our approach, we use manifold alignment with parallel fields to model disparate media $\mathcal{A}$ and $\mathcal{B}$ data and project their relationship into intermediate spaces, and the semantic correlations are determined during the process of manifold alignment. In this way, the cross media retrieval task can be resolved by the determined semantic correlation. We next show our proposed method, the parallel field alignment retrieval (PFAR), in detail.

3.2 Parallel field alignment retrieval

In order to learn intermediate spaces $\mathcal{I}^{\mathcal{A}}$ and $\mathcal{I}^{\mathcal{B}}$, An optimal correspondence between the representations in the original spaces $\mathcal{R}^{\mathcal{A}}$ and $\mathcal{R}^{\mathcal{B}}$ [24] is needed. One possible way is to apply the subspace learning framework which utilize some extremely well developed dimensionality reduction approaches, such as Principal Component Analysis (PCA) [13] or Latent Semantic Indexing (LSI) [6].

Our approach here utilizes manifold alignment algorithms [10, 30] to find mappings defined on the media data manifolds, and use these mappings to align the underlying media manifold representations in intermediate semantic spaces. In the sense of manifold alignment, the semantic correlations are learned after the alignment of the underlying media manifold representations. In addition, we would like to use parallel fields, which preserve the metric of the manifold to measure the disparate media data manifolds. Most of existing manifold alignment algorithms focus on ensuring the first order smoothness of the function. However, researchers have shown that the second order smoothness of the function is particularly important for preserving the metric. And the second order smoothness of the function is equivalent to the parallelism of the gradient field of the function [18]. Thus, we propose to learn two mapping functions and two vector fields simultaneously with two constraints in the process of manifold alignment. By designing the two constraints, we ensure each vector field to be close to the gradient field of each corresponding mapping function, and the vector fields should be as parallel as possible. In this way, the second order smoothness of the function is also ensured. The whole concept of our proposed parallel field alignment retrieval approach is shown in Figure 4.

We now consider aligning disparate media data manifolds, given some additional information of datasets, from the parallel field perspective. In our approach, the desired coordinates for labeled samples are provided. In detail, the coordinates of labels indicate the intrinsic information of paired data samples within the manifolds. With a small number of labeled examples, it is crucial to exploit manifold structures of datasets in manifold alignment [10]. Specifically, we first estimates the gradient field of the prediction function by a vector field, and then require the vector field to be as parallel as possible.

In the setting of parallel field alignment retrieval (PFAR), we denote the notations $A \subset \mathcal{R}^{\mathcal{A}}$ and $B \subset \mathcal{R}^{\mathcal{B}}$ as the datasets, and $\mathbf{s} = \{s_1, s_2, \ldots, s_l\}$ and $\mathbf{t} = \{t_1, t_2, \ldots, t_l\}$ are the vectors referred to the target paired information of the l "labeled" data samples for A and B respectively. Let f and g denote two mapping functions defined on the data manifolds that match known target labeled pairs. Following the above analysis, we try to integrate vector fields V_A and V_B on the manifold with two constraints to our alignment functions:

$$E(f, V_A) = \frac{1}{l_A} \sum_{i=1}^{l_A} (f(a_i) - s_i)^2 + \mu_{A,1} R_1(f, V_A) + \mu_{A,2} R_2(V_A) \tag{9}$$

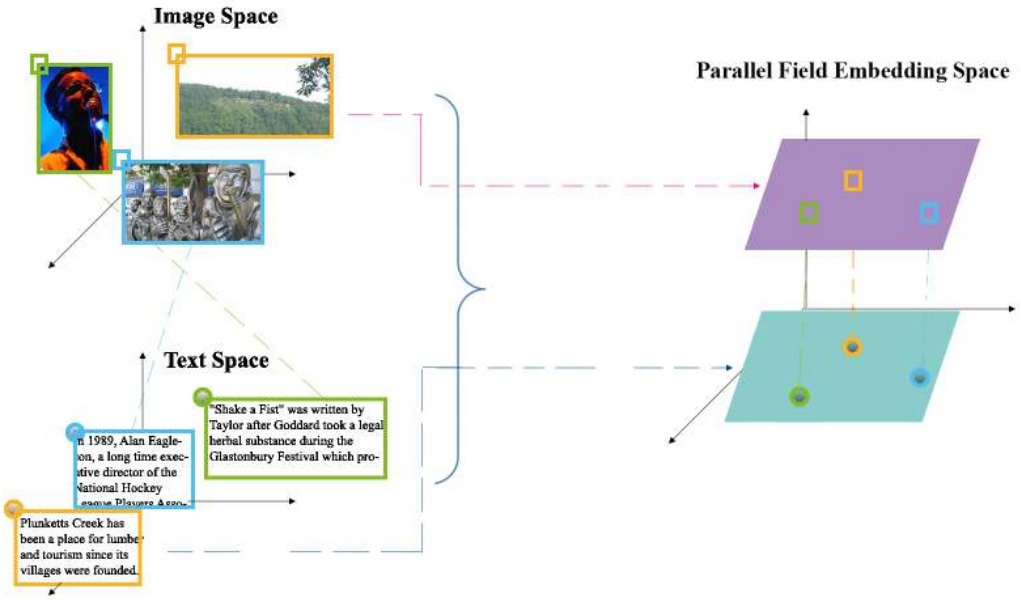


Figure 4: The PFAR concept illustration.

and

$$E(g, V_B) = \frac{1}{l_B}\sum_{i=1}^{l_B}(g(b_i)-t_i)^2+\mu_{B,1}R_1(g, V_B)+\mu_{B,2}R_2(V_B), \tag{10}$$

where $a_i \in A$ and $b_i \in B$, R_1 and R_2 are regularizers defined in Equation 3 and 4, R_1 enforces vector fields V_A and V_B to be close to gradient fields ∇f and ∇g of mapping functions. As shown in Section 2, if R_2 vanishes, V_A and V_B must be parallel fields.

In order to explore how underlying data manifolds can be mapped into intermediate spaces and then aligned to each other through a common set of paired media data information in the process of PFAR, we should make sure that the vector field is close to the gradient field of the function and further the vector field should be as parallel as possible.

Next we show that how to discretize the continuous objective function. First of all, we give some notations:

- Let $T_{a_i}\mathcal{M}$ denote the d dimensional tangent space of a_i on the manifold $\mathcal{M}$.
- Let $T_i \in \mathbb{R}^{m\times d}$ denote the matrix whose columns constitute an orthonormal basis for $T_{a_i}\mathcal{M}$.
- Let V_{a_i} denote the value of the vector field V at data point a_i.

Following the above notations, define $P_i = T_i T_i^T$. It can be shown that P_i is the unique orthogonal projection from $\mathbb{R}^m$ onto the tangent space $T_{a_i}\mathcal{M}$ [8]. According to the definition of vector fields, each vector V_{a_i} should be on the tangent space $T_{a_i}\mathcal{M}$. Thus we can represent V_{a_i} by the coordinates of tangent spaces, *i.e.*, $V_{a_i} = T_i v_i$. Let $\mathbb{V}_A = (v_1^T, \ldots, v_n^T)^T \in \mathbb{R}^{dn}$ be a dn-dimension column vector which concatenates all the v_i's, $i \in (1, \ldots, n)$.

Then the regularizers R_1 and R_2 can be discretely reduced to:

$$R_1(f, \mathbb{V}_A) = \sum_i \sum_{j\sim i} w_{ij}((a_i - a_j)^T T_i v_i - f_j + f_i)^2 \tag{11}$$

and

$$R_2(\mathbb{V}_A) = \sum_i \sum_{j\sim i} w_{ij}\|P_i T_j v_j - T_i v_i\|^2, \tag{12}$$

where w_{ij}, weight parameters, which can be approximated by heat kernel weights $\exp(-\|a_i - a_j\|^2/\delta)$ or by 0 - 1 weights for simplicity.

Now, let $\mathbb{I}_A$ denote an $n\times n$ diagonal matrix where $\mathbb{I}_{Aii} = 1$ if a_i is labeled and $\mathbb{I}_{Aii} = 0$ otherwise. Then the discrete form of our parallel field alignment objective function $E(f, V_A)$ can be written as:

$$\begin{aligned} E(f, \mathbb{V}_A) = &\frac{1}{l_A}(f - \mathbf{s})^T \mathbb{I}_A (f - \mathbf{s}) \\ &+ \mu_{A,1}\sum_i \sum_{j\sim i} w_{ij}((a_i - a_j)^T T_i v_i - f_j + f_i)^2 \\ &+ \mu_{A,2}\sum_i \sum_{j\sim i} w_{ij}\|P_i T_j v_j - T_i v_i\|^2 \end{aligned}. \tag{13}$$

The optimal solution to this objective function is then obtained via solving the following linear systems:

$$\begin{pmatrix} \frac{1}{l_A}\mathbb{I}_A + 2\mu_{A,1}L_A & -\mu_{A,1}C_A^T \\ -\mu_{A,1}C_A & \mu_{A,1}G_A + \mu_{A,2}K_A \end{pmatrix} \begin{pmatrix} f \\ \mathbb{V}_A \end{pmatrix} = \begin{pmatrix} \frac{1}{l_A}\mathbf{s} \\ 0 \end{pmatrix}, \tag{14}$$

where L denotes the Laplacian matrix of the graph with weights w_{ij}, G_A is a $dn \times dn$ block diagonal matrix, and C_A is a $dn \times dn$ block matrix. Let us denote G_{Aii} and C_{Ai} as the i-th $d \times d$ diagonal block of G and the i-th $d \times n$ block of C respectively, and $z_{ij} \in \mathbb{R}^n$ is a selection vector of all zero

elements except for the i-th element being -1 and the j-th element being 1. Then G_{Aii} and C_{Ai} are defined as

$$G_{Aii} = \sum_{j\sim i} w_{ij} T_i^T (a_j - a_i)(a_j - a_i)^T T_i \tag{15}$$

and

$$C_{Ai} = \sum_{j\sim i} w_{ij} T_i^T (a_j - a_i) z_{ij}^T, \tag{16}$$

K_A is a $dn \times dn$ sparse block matrix. If we use $K_{A_{ij}}$ to index each $d \times d$ block in A, $i, j \in (1, \ldots, n)$, we have

$$K_{A_{ii}} = \sum_{j\sim i} w_{ij}(Q_{ij} Q_{ij}^T + I) \tag{17}$$

and

$$K_{A_{ij}} = \begin{cases} -2w_{ij} Q_{ij}, & \text{if } a_i \sim a_j \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

where $Q_{ij} = T_i^T T_j$.

Similarly, the optimal solution for data set B is as follows:

$$\begin{pmatrix} \frac{1}{l_B}\mathbb{I}_B + 2\mu_{B,1} L_B & -\mu_{B,1} C_B^T \\ -\mu_{B,1} C_B & \mu_{B,1} G_B + \mu_{B,2} K_B \end{pmatrix} \begin{pmatrix} g \\ \mathbb{V}_B \end{pmatrix} = \begin{pmatrix} \frac{1}{l_B}\mathbf{t} \\ 0 \end{pmatrix}. \tag{19}$$

Given two datasets A and B with target paired values of labeled data samples, the solutions to f and g of Equation 14 and Equation 19 can be used to estimate coordinates of the other unlabeled data points in intermediate spaces, which further can be utilized to align their intrinsic data manifolds.

Given any query q from A, we use the mapping function f obtained in the training step to project the query q into the intermediate space. And the projected q is then semantic correlated aligned in the intermediate space. To find the best match samples in B, we can use the metric defined as: let $F = (f_1, \ldots, f_r)^T$ and $G = (g_1, \ldots, g_r)^T$ be the r dimensional representations of aligned manifolds of A and B. If the coordinates in F and G are aligned from known coordinates, the distance between $a_i \in A$ and $b_j \in B$ is then given by [10]:

$$d(a_i, b_j)^2 = \sum_k |F_{ik} - G_{jk}|^2, \tag{20}$$

then the best match $b_j \in B$ to $a \in A$ is given by:

$$\arg\min_j \; d(a, b_j). \tag{21}$$

The alignment framework is similar to some of the existing alignment methods [9, 10, 30]. However, most of the existing alignment approaches focus on preserving the pairwise similarity between data pairs. In this case, they may not preserve the relative order of the similarity measure. For example, suppose object A and object B are similar, object A and object C are also similar, however B is more similar to A than C. In existing alignment methods, they can find a space in which A, B, C are close, but they cannot tell whether B is closer to A than C or not. By using vector fields, we require the mapping function varies linearly along the geodesics on the manifolds and naturally the order can be preserved. This property is extremely important for multimedia retrieval problems.

4. EXPERIMENTS

In this section, we conduct some extensive experimental evaluation to demonstrate the effectiveness of our proposed *Parallel Field Alignment Retrieval* (PFAR) approach.

4.1 Dataset

The evaluation of a cross media retrieval system usually needs a document corpus with paired contents from disparate domains of multimedia source. In this experiment, we use the recently proposed Wikipedia dataset composed of text and image pairs[1] [24]. This real world dataset consists of a continually updated collection of Wikipedia's featured articles spread over 29 categories, since 2009. The articles are accompanied by one or more images from the Wikipedia Commons, supplying a pairing of the desired kind. Since some categories are very scarce, we will only consider the top 10 most categories in the experiment.

Each article was split into sections based on its section headings, and each image was assigned to the corresponding section in which it was placed by the author(s) of the Wikipedia page. Then this dataset was pruned to keep only sections that contained a single image and at least 70 words. The final corpus contains 2866 text-image pairs, each belongs to one of 10 semantic categories. And the corpus is random split into a training set with 2173 documents, and a test set with 693 documents [24]. The detail information of this corpus is summarized in Table 1.

Table 1: Experiment Dataset Summary [24]

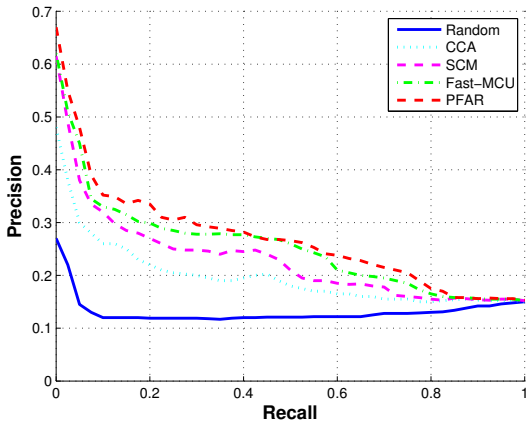
Category	Training	Query	Total
Art & architecture	138	34	172
Biology	272	88	360
Geography & places	244	96	340
History	248	85	333
Literature & theatre	202	65	267
Media	178	58	236
Music	186	51	237
Royalty & nobility	144	41	185
Sport & recreation	214	71	285
Warfare	347	104	451
Summary	2173	693	2866

4.2 Data representation and evaluation

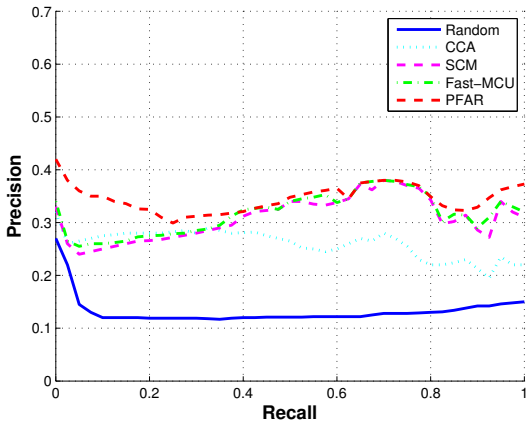
The data representation is similar to those of two previous works [20, 24]. In terms of image representations $\mathcal{R}^I$, we represented the image using the histogram over the popular scale-invariant feature transformation (SIFT) [19] with the codebook of 512 codewords. The text representation is derived from a Latent Dirichlet Allocation (LDA) [1] model with 10 topics (for training data, label information is derived based on the topics). Thus, in $\mathcal{R}^T$, text documents are represented by their topic assignment probability distributions among these 10 topics.

To test the proposed PFAR approach on real data, we conduct the retrieval task with two parts: text retrieval using an image query, and image retrieval using a text query. In the first part, each image in the test set is used as a query, and the goal is to rank all the texts in the test set based on their match to the query image. In the second, a text query

[1]http://www.svcl.ucsd.edu/projects/crossmodal



(a) PR curve of text query task



(b) PR curve of image query task

Figure 5: Precision recall curves: (a) text query and (b) image query

is used to rank the images. In both parts, the performance is measured using precision-recall (PR) curves and the mean average precision (MAP).

In the experiment, our approach learns the aligned intermediate spaces by incorporating the information of media data. We use $k = 8$ nearest neighbors method to construct the neighborhood for each media source in formulation of Equation 9, and we apply the 10-fold cross validation to select the parameters (*e.g.*, $\mu_{A,1}$ for R_1, $\mu_{A,2}$ for R_2). To demonstrate the performance of PFAR, we compare a number of state-of-the-art cross media retrieval approaches, CCA [12], Semantic Correlation Matching (SCM) [24], Fast version of Maximum Covariance Unfolding (Fast-MCU) [20] and Manifold Alignment (MA) [10] with PFAR.

In these retrieval approaches, given a test sample (image or text), it is first projected into the learned intermediate space. For CCA and SCM, this involves a linear transformation to the low dimensional subspace, while for Fast-MCU, the nearest neighbors of the test point among the training samples in the original space are used to obtain a mapping of the point as a weighted combination of these neighbors [20].

For PFAR, we use training datasets to learn the intermediate spaces during the parallel field alignment process and two manifolds are aligned at the same time. The same mapping is then applied to the projection of the neighbors in the learned intermediate space to compute the projection of the test point. To perform retrieval, all the test samples from both models, image and text, are projected on the aligned manifolds, respectively. For a given test point from one kind of media, we use the correlation distance shown in Equation 21 to compute its distance to all the other projected test points of the other medium, and then these distances are ranked. If a retrieved sample belongs to the same category as the query, it is considered to be correct.

4.3 Test of the cross media retrieval

The result of the retrieval task is shown in Table 2, which summarizes the MAP scores obtained for the 5 cross media retrieval approaches. This table contains scores for both image retrieval from a text query, and text retrieval from an image query, and the average. The performance of the random retrieval test is also shown to indicate the baseline chance level. From Table 2, it is clear that our proposed approach PFAR outperforms CCA, SCM and Fast-MCU in both image-to-text (image query) and text-to-image (text query) cross media retrieval tasks. PFRA leads non-trivial improvements over other approaches. In both parts, the average MAP of our approach is more than 2.5 times that of random method.

Table 2: Retrieval Performance (MAP Scores)

Experiments	Image Query	Text Query	Average
Random	0.118	0.118	0.118
CCA	0.246	0.196	0.221
SCM	0.274	0.225	0.250
Fast-MCU	0.287	0.224	0.256
MA	0.262	0.225	0.243
PFAR	**0.298**	**0.273**	**0.286**

To further analyze the performance of our proposed approach, we also presents the results of precision-recall (PR) curves for both image and text queries in Figure 5. According to the results of Figure 5, it is clear that our proposed approach PFAR, Fast-MCU, SCM, and CCA all gain improvements over random retrieval at all levels of recall. Compared the PR curves of PFAR with those of SCM and Fast-MCU, it shows that PFAR gets higher precision at all levels of recall for both text queries (on the left) and image queries (on the right). It shows that the PFAR approach is more effective to precisely find matches than the other approaches.

Figure 6 illustrates two examples of text queries and the top ranked images retrieved by PFAR to provide the visualization of cross media retrieval. These two examples are chosen from geography category and sport category, respectively. In each case, the query text is shown at the top, and the first image in the images row is the groundtruth image. As indicated in Figure 6, top four retrieved images are shown next to the groundtruth image. We can see that the retrieved images are quite semantic correlated to the query texts.

Figure 7 shows an example about topics distribution of retrieved texts by CCA, SCM, Fast-MCU, MA and our pro-

Second growth forests have since covered most of the clear-cut land. The beginnings of today's protected areas were established in the late nineteenth and early twentieth centuries: Pennsylvania's state legislature authorized the acquisition of abandoned clear-cut land in 1897, creating the state forest system. The Game Commission began acquiring property for State Game Lands in 1920, and established the Northcentral State Game Farm on Plunketts Creek in 1945 to raise wild turkey. It was converted to ringneck pheasant production in 1981, and, as of 2007, it is one of four Pennsylvania state game farms producing about 200,000 pheasants each year for release on land open to public hunting. The Northcentral State Game Farm is in the Plunketts Creek valley just south of Proctor, and a part of it is on the right bank of Loyalsock Creek downstream of the confluence. The Loyalsock State Game Farm is 13 miles (21 km) downstream on Loyalsock Creek, at the village of Loyalsockville. When a May 2007 fire destroyed a brooder house there just days before 18,000 pheasant chicks were due to hatch, the eggs were transferred to the nearby Northcentral State Game Farm without reduction in the production goal.

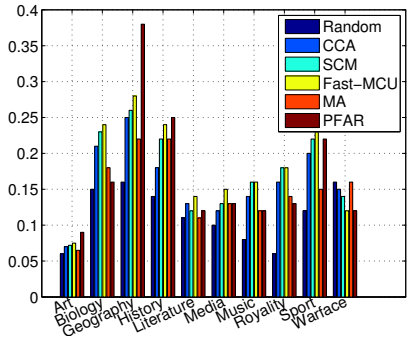
(a) Text query from geography category

On March 30, 1993, it was announced that Gil Stein, who at the time was the president of the National Hockey League, would be inducted into the Hall of Fame. There were immediate allegations that he had engineered his election through manipulation of the hall's board of directors. Due to these allegations, NHL commissioner Gary Bettman hired two independent lawyers, Arnold Burns and Yves Fortier, to lead an investigation. They concluded that Stein had "improperly manipulated the process" and "created the false appearance and illusion" that his nomination was the idea of Bruce McNall. They concluded that Stein pressured McNall to nominate him and had refused to withdraw his nomination when asked to do so by Bettman. There was a dispute over McNall's role and Stein was "categorical in stating that the idea was Mr. McNall's." They recommended that Stein's selection be overturned, but it was revealed Stein had decided to turn down the induction before their announcement.
In 1989, Alan Eagleson, a long time executive director of the National Hockey League Players Association, was inducted as a builder. He resigned nine years later from the Hall after pleading guilty to mail fraud and embezzling hundreds of thousands of dollars from the NHL Players Association pension funds."Honoured members: the Hockey Hall of Fame", p. 167 His resignation came six days before a vote was scheduled to determine if he should be expelled from the Hall. Originally, the Hall of Fame was not going to become involved in the issue, but was forced to act when dozens of inductees, including Bobby Orr, Ted Lindsay and Brad Park, campaigned for Eagleson's expulsion, even threatening to renounce their membership if he was not removed. He became the first member of a sports hall of fame in North America to resign.

(b) Text query from sport category

Figure 6: Two examples of text queries and the top images retrieved by PFAR. The query texts of (a) and (b) are chosen from geography category and sport category, respectively. In both (a) and (b), first image in bottom row is the groundtruth image, and others are top ranked retrieved images.

(a) Query image



(b) Topics distribution of retrieved texts

Topography [edit]

See also: List of Seattle parks, Bodies of water of Seattle, and Regrading in Seattle

Seattle is located between the saltwater Puget Sound (an arm of the Pacific Ocean) to the west. The city's chief harbor, Elliott Bay, is part of Puget Sound, which makes the city an o Puget Sound, are the Kitsap Peninsula and Olympic Mountains on the Olympic Peninsula; Washington and the eastside suburbs, are Lake Sammamish and the Cascade Range. Lal Puget Sound through the Lake Washington Ship Canal (consisting of two man-made canal Chittenden Locks at Salmon Bay, ending in Shilshole Bay on Puget Sound).

The sea, rivers, forests, lakes, and fields surrounding Seattle were once rich enough to su sedentary hunter gatherer societies. The surrounding area lends itself well to sailing, skiing year round.[24][25]

The city itself is hilly, though not uniformly so.[9] Like Rome, the city is said to lie on sever include Capitol Hill, First Hill, West Seattle, Beacon Hill, Queen Anne, Magnolia, and the fo Mount Baker, and Crown Hill neighborhoods are technically located on hills as well. Many c center, with Capitol Hill, First Hill, and Beacon Hill collectively constituting something of a r Elliott Bay and Lake Washington.[26] The break in the ridge between First Hill and Beacon I of the many regrading projects that reshaped the topography of the city center.[27] The top construction of a seawall and the artificial Harbor Island (completed 1909) at the mouth of Green River. The highest point within city limits is at High Point in West Seattle, roughly lo include Crown Hill, View Ridge/Wedgwood/Bryant, Maple Leaf, Phinney Ridge, Mt. Baker

(c) The groundtruth text: Seattle from Wikipedia

Figure 7: A comparison example of cross media retrieval with given image query.

posed PFAR. Given a query image, we compared the topics distribution of all retrieved texts with respect to above methods. In Figure 7, the query image is taken from Seattle, the groudtruth text gives some description about the geography of Seattle. The result of PFAR can accurately give the topics distributions over other methods.

Actually, all methods tend to retrieve texts that are related specifically to the query image. However, PFAR is able to retrieve texts which are closer to the query image on the categorical level. This also indicates that the abstraction work with relative order infomation preserved by PFAR is especially important for exploratory tasks.

According to the experiment result analysis, PFAR can be viewed as one kind of semantic correlation spaces projection methods. In addition, the parallel field in the alignment process, which preserves the geodesic distance of media manifolds, also significantly contribute to the performance of the text and image cross media retrieval.

5. CONCLUSIONS

In this paper, we have proposed a novel approach for cross media retrieval, called Parallel Field Alignment Retrieval (PFAR), which integrates a manifold alignment framework from the perspective of vector fields. Our goal is to solve the fundamental problem in cross media retrieval area. Given a query $Q_{\mathcal{A}} \in \mathcal{R}^{\mathcal{A}}$ (or $Q_{\mathcal{B}} \in \mathcal{R}^{\mathcal{B}}$), the goal of cross media retrieval is to return the closest matches in the $\mathcal{B}$ (or $\mathcal{A}$) media space $\mathcal{R}^{\mathcal{B}}$ (or $\mathcal{R}^{\mathcal{A}}$).

Due to the semantic gap between low level features and high level semantic concepts of multimedia data, it is not easy for us to reveal the underlying semantic correlations among the heterogeneous multimedia data. Previous studies have shown that manifold alignment method is an appropriate model for pair matching between heterogeneous data sources. Hence, we consider the cross media retrieval as a manifold alignment problem and use parallel fields, which can effectively preserve the metric of media data manifolds, to model heterogeneous multimedia data and project their relationship into intermediate latent semantic spaces during the process of manifold alignment. And the most important factor is that the semantic correlations are also determined in the manifold alignment process.

Finally, the experimental results from a real world data illustrate the validity of our approach. In text and image cross media retrieval tasks, our approach attains significant in the retrieval performance over state-of-the-art cross media retrieval methods.

6. ACKNOWLEDGMENTS

This work is supported by National Basic Research Program of China (973 Program) under Grant 2012CB316400 and National Natural Science Foundation of China (Grant No: 61222207, 61125203, 61233011, 90920303). Mao's and Pei's research is supported in part by an NSERC Discovery Grant and a BCFRST NRAS Endowment Research Team Program project. All opinions, findings, conclusions and recommendations in this paper are those of the author and do not necessarily reflect the views of the funding agencies.

7. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.

[3] C. Cotsaces, N. Nikolaidis, and I. Pitas. Video shot detection and condensed representation. a review. *Signal Processing Magazine, IEEE*, 23(2):28–37, 2006.

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

[5] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):5:1–5:60, 2008.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[7] H. Escalante, C. Hérnadez, L. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 172–179. ACM, 2008.

[8] G. H. Golub and C. F. V. Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[9] J. Ham, D. D. Lee, and L. K. Saul. Learning high dimensional correspondences from low dimensional manifolds. In *Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 34–41, 2003.

[10] J. Ham, D. D. Lee, and L. K. Saul. Semisupervised alignment of manifolds. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence*, volume 10, pages 120–127, 2005.

[11] X. He, W. Ma, and H. Zhang. Learning an image manifold for retrieval. In *Proceedings of the 12th ACM International Conference on Multimedia*, pages 17–23. ACM, 2004.

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

[13] I. Jolliffe. *Principal component analysis*, volume 2. Wiley Online Library, 2002.

[14] T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo. Combining image captions and visual analysis for image concept classification. In *Proceedings of the 9th International Workshop on Multimedia Data Mining*, pages 8–17. ACM, 2008.

[15] J. Lafferty and L. Wasserman. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems 20*, pages 801–808, Cambridge, MA, 2008.

[16] M. Lestari Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: overview of the imageclefphoto task 2009. pages 45–59. Springer, 2010.

[17] B. Lin, X. He, C. Zhang, and M. Ji. Parallel vector field embedding. *The Journal of Machine Learning Research*, 2013.

[18] B. Lin, C. Zhang, and X. He. Semi-supervised regression via parallel field regularization. In *Advances in Neural Information Processing Systems 24*, pages 433–441, Cambridge, MA, 2011.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[20] V. Mahadevan, C.-W. Wong, T. T. Liu, N. Vasconcelos, and L. K. Saul. Maximum covariance unfolding: Manifold learning for bimodal data. In *Advances in Neural Information Processing Systems 24*, pages 918–926, 2011.

[21] B. Nadler, N. Srebro, and X. Zhou. Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *Advances in Neural Information Processing Systems 22*, pages 1330–1338, 2009.

[22] P. Petersen. *Riemannian Geometry*, volume 171. Springer, 2006.

[23] T. Pham, N. Maillot, J. Lim, and J. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 439–444. ACM, 2007.

[24] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. 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.

[25] N. Rasiwasia, P. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia*, 9(5):923–938, 2007.

[26] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[27] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th ACM International Conference on Multimedia*, pages 399–402. ACM, 2005.

[28] T. Tsikrika and J. Kludas. Overview of the wikipediamm task at imageclef 2008. pages 539–550. Springer, 2009.

[29] C. Wang and S. Mahadevan. Manifold alignment using procrustes analysis. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1120–1127. ACM, 2008.

[30] C. Wang and S. Mahadevan. A general framework for manifold alignment. 2009.

[31] C. Wang and S. Mahadevan. Heterogeneous domain adaptation using manifold alignment. volume 2, pages 1541–1546. AAAI Press, 2011.

[32] G. Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1374. IEEE, 2009.

[33] T. Westerveld. Probabilistic multimedia retrieval. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 437–438. ACM, 2002.

[34] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 175–184. ACM, 2009.

[35] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 10(3):437–446, 2008.

[36] Y. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, 10(2):221–229, 2008.