Learning Semantic Distance from Community-Tagged Media Collection

[†]Guo-Jun Qi, [‡]Xian-Sheng Hua, and [‡]Hong-Jiang Zhang
[†]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
qi4@illinois.edu
[‡]Microsoft Corporation
{xshua, hjzhang}@microsoft.com

ABSTRACT

This paper proposes a novel semantic-aware distance metric for images by mining multimedia data on the Internet, in particular, web images and their associated tags. As well known, a proper distance metric between images is a key ingredient in many realistic web image retrieval engines, as well many image understanding techniques. In this paper, we attempt to mine a novel distance metric from the web images by integrating their visual content as well as the associated user tags. Different from many existing distance metric learning algorithms which utilize the dissimilar or similar information between images pixels or features in signal level, the proposed scheme also takes the associated user-input tags into consideration. The visual content of images is also leveraged to respect an intuitive assumption that the visual similar images ought to have a smaller distance. A semi-definite programming is formulated to encode the above two aspects of criteria to learn the distance metric and we show such an optimization problem can be efficiently solved with a closed-form solution. We evaluate the proposed algorithm on two datasets. One is the benchmark Corel dataset and the other is a real-world dataset crawled from the image sharing website Flickr. By comparison with other existing distance learning algorithms, competitive results are obtained by the proposed algorithm in experiments.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms

Algorithms, Theory, Experimentation

Keywords

Distance Metric Learning, Web Image Search

MM'09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10 ...$10.00.

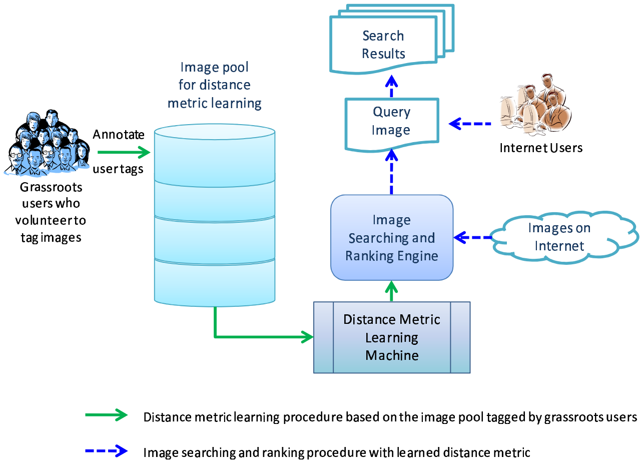

Search Results
Image pool for distance metric learning
Query Image
Internet Users
Annotate
user tags
Grassroots users who volunteer to tag images
Image Searching and Ranking Engine
Images on Internet
Distance Metric Learning Machine
Distance metric learning procedure based on the image pool tagged by grassroots users
Image searching and ranking procedure with learned distance metric

Figure 1: The distance metric is mined from the visual content of images together with the tags annotated by the grassroots users. The learned distance metric then is applied to retrieve images on the Internet by ranking their relevances to the query.

1. INTRODUCTION

Many image-sharing websites, such as Flickr and Corbis, have emerged which significantly promote the storage, sharing, exchange and propagation of images among users. Meanwhile, based on these infrastructures of image-sharing social networks, amount of grassroots users voluntarily provide their tags to annotate these images everyday. Consequently, many research opportunities arise to leverage these user tags to boost image search and retrieval. By exploring these free tags given by users, we believe a user-driven image search system, which is more consistent with users' subjectivity, can be built since these tags contain the user intentions and their fashion taste in time. In particular, we propose in this paper to mine a distance metric from these web images and their associated user tags, which can be directly applied to retrieve images on the web.

As well known, an appropriate distance metric plays a key role in many image search systems as a fundamental tool to measure the relevance between different images across the website [7] [17]. Images in collection can be retrieved and ranked by their distances to the query image given by user (See Figure 1 for example). The smaller distance between the query and the image means a higher relevance between

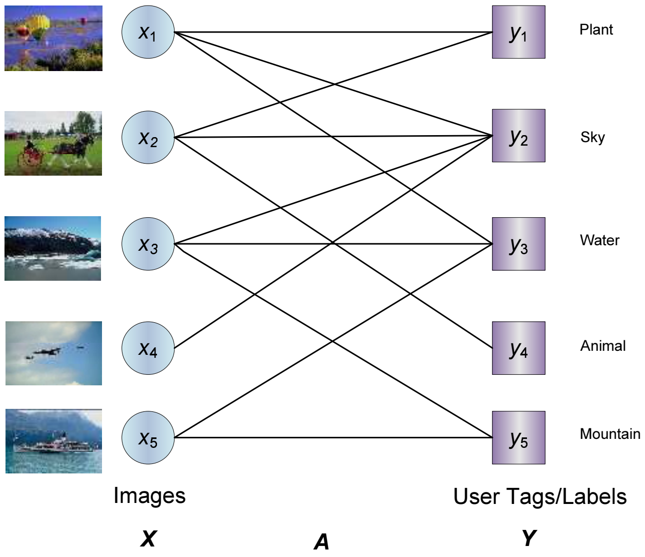


Figure 2: Some images and their associated user tags/labels represented by a bipartite graph.

them. On the other hand, the learned distance metric can also be applied to automatically annotate web images with the tags of the nearest image in the distance space according to the annotation-by-search paradigm [19][20].

Up to now, many distance learning algorithms have been proposed to reveal the intrinsic distance metric by exploiting various information contained in the training set. For example, the annotated similar or dissimilar sample pairs in the training set can be used as side-information to learn the distance metric [23] [5]. In the other formulation, the distance can otherwise be learned from "chunklets" of similar samples instead of the similar pairs, which is the main idea of Relevance Component Analysis (RCA) algorithm [1]. Furthermore, Schultz et al. [14] proposed to learn the distance function from the relative comparisons which are believed to be more easily obtained in many real-world settings.

The above mentioned methods have gained significant success in many image retrieval applications [17] [16] [24]. However, all of them are aiming at learning the distance functions defined over a single input space with homogeneous data objects, which can be represented either as a set of feature vectors with the same dimension or a homogenous graph with nodes of a single type. On the contrary, for image retrieval application in the website (e.g., Flickr and Yahoo! Image) environment, each image is often annotated with many customized *tags* by users in addition to their own *visual* features. These two different types of data interact with each other in these dual spaces (i.e., the image visual feature space and tag space). Obviously, the existing homogeneous distance learning algorithm cannot be directly adopted to mine a proper distance metric from these abundant user tags which interact with the images on the websites. In other words, these interactions between images and tags are characterized by tags being annotated to images or not. Such a heterogeneous data structure and their interactions can be represented by a bipartite graph with two types of samples (i.e., the image samples and tag samples here) as illustrated Figure 2. So we wonder if there exists a direct method to mine an image semantic distance from these interactions, which is the main focus of this paper. This motivates us to learn the distance functions over the above heterogeneous data structure with the interacting images/tags relations. The learned semantic distance should reflect the relevance between images in the semantic meanings contained in the user tags. Note that most images on the Internet have not been tagged by users on most common websites except for those on photo-sharing websites. Therefore, with the learned semantic distance from the community-tagged images, we can apply the distance to retrieve large amount of images with no user tags. For example, for a given image query, the image search engine can rank the retrieved images by relevance in ascending order of the semantic distances from these retrieved ones to the query.

On the other hand, it is worth noting that the user tags on the websites suffer from significant noise, such as random tagging, misspelling, and personalized ambiguity tags[3][22]. To reduce these noisy effect, we propose a visual regularizer based on the content information of the images. Using the visual features, the regularizer uses the image similarity to reduce the overfitting risk of noisy tags. That is to say, a pair of visually similar images should have smaller semantic distance and vice versa. By combining the semantic information in the user tags and the visual content of images, a more robust semantic distance can be expected in this paper.

In the bipartite graph structure illustrated in Figure 2, each image can be assigned by more than one tag/label. This is quite different from many other distance metric learning algorithms, such as RCA [1], in which each image is exclusively assigned into one chunklet and only the images in the same chunklet are considered to be similar. The proposed distance learning method based on the above bipartite graph removes this one-chunklet (class) restriction and each image can be annotated by more than one labels/tags. This multi-label setting [11][12] puts more challenges on learning the distance metric because the images cannot be simply considered to be similar or dissimilar like in RCA formulation. Accordingly, we call the proposed distance learning method by Multi-Label Distance Metric Learning (ML-DML) for convenience.

Although there are many existing literatures [6] [13] [10] to study co-clustering problem in bipartite graph as Figure 2, to the best of our knowledge no effort has been made on learning the distance on such a heterogeneous data structure. It is true that co-clustering can provide significant information on the underlying knowledge of data structures in different data types, but it also suffers from quite a few problems.

1. The obtained clusters cannot be directly applied to the image retrieval. Although these clusters provide important structures on the training set (e.g., the user community and video groups sharing a common topic), there is no direct way to apply them to retrieve the images. Moreover in many learning algorithms, the clustering results cannot be directly embedded into their formulations, such as k-nearest neighborhood and kernel-based methods. In contrast, it is direct to use the learned distance functions to obtain a list of rank images ordered by their relevance to the query image.

2. Most existing co-clustering algorithms based on bipartite

graph only utilize the interactions between images and user tags but ignore the visual content of the images themselves. For example, Dhillon et al. [6] uses the relations between two heterogeneous data to construct a similarity matrix, based on which the spectral clustering algorithm is applied to cluster these heterogeneous samples. However, it is obvious that both the visual content of images and user tags provide valuable information to explore the underlying the structure of these heterogenous samples. For example, if an image is more visually similar to the query image, it should be more relevant to this query.

3. The co-clustering methods only focus on the samples contained in training set. It is often difficult (if not impossible) to generalize the obtained cluster structures to unseen data. Instead the distance learning attempts to learn distance functions defined over the whole input spaces so that the generalization to the unseen data does not bring any extra effort.

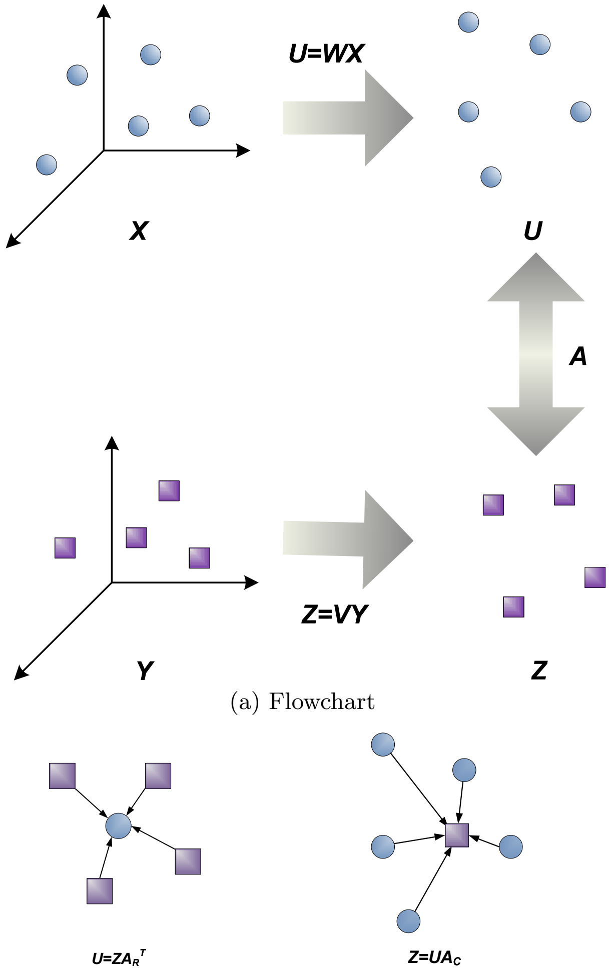
In a brief summary, a multi-label distance learning algorithm is proposed in this paper to address how to learn the distance metric from user tags. To be detailed in Section 2, the learned distance metric utilizes the interacting relationship between the images and user tags, as well as the visual image content. A semi-definite optimization problem is formulated to learn such a multi-label distance and a closed-formed optimum can be derived from this optimization problem. We further extend the solution to nonlinear one by incorporating the kernelization into the formulation. In Section 3, we will uncover the underlying connection between the proposed semantic distance and the RCA-based distance function. It is proved that the latter one is a special case of the proposed multi-label distance in this paper. We conduct the experiments in Section 4 to evaluate the proposed algorithm by comparing with other widely-used distances. Finally, we conclude in Section 5.

2. MINING MULTI-LABEL DISTANCE METRIC

In this section, we first define some notations and problem setting for multi-label distance metric learning on bipartite graph in Section 2.1. The proposed algorithm is formulated in Section 2.2, followed by an extension into a nonlinear version in Section 2.3.

2.1 Notation Definition and Problem Setting

Assume a bipartite-graph has two different types of samples, and we represent them by two matrices $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$ and $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_m]$ where each column represents one sample. Here $\boldsymbol{x}_i \in \mathcal{R}^{d_X}$ is the ith sample of the first type and $\boldsymbol{y}_j \in \mathcal{R}^{d_Y}$ is the jth sample of the second type; d_X and d_Y are the dimensions of feature spaces of these two data types. The relation matrix $A = (a_{ij})_{n \times m}$ denotes the interactions between these two data types. In particular, for image retrieval problem, a_{ij} is used to indicate a tag $\boldsymbol{y}_j$ is assigned to the image $\boldsymbol{x}_i$ by the users if $a_{ij} = 1$, or not if $a_{ij} = 0$; a_{ij} can also be used to denote the times that a user tag assigned to an image. The goal aims at computing two distance functions $dist(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ and $dist(\boldsymbol{y}, \tilde{\boldsymbol{y}})$ defined on these two input feature spaces $\mathcal{R}^{d_X}$ and $\mathcal{R}^{d_X}$. In the following sections, we derive $dist(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ in detail. $dist(\boldsymbol{y}, \tilde{\boldsymbol{y}})$ can be easily derived in the same manner.



(a) Flowchart

(b) Reconstruct U from Z (c) Reconstruct Z from U

Figure 3: The flowchart of the proposed multi-label distance metric learning algorithm. In subfigure (a), two types of samples X and Y are first transformed into two latent spaces U and Z, where U and Z are correlated by the relational matrix A. In subfigures (b) and (c), U and Z can then be reconstructed by row-normalized and column-normalized relational matrices, respectively.

2.2 Main Idea and Solution

First we transform two types of samples into two latent spaces with the same dimension by two linear transformations matrices W and V, respectively (see Figure 3 for an illustration). Here the transformed samples are denoted by $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots \boldsymbol{u}_n]$ and $Z = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots \boldsymbol{z}_m]$ where $\boldsymbol{u}_i = W\boldsymbol{x}_i$ and $\boldsymbol{z}_j = V\boldsymbol{y}_j$. The two distance functions can then be computed in the transformed latent spaces.

To derive the transformation matrices, we expect in these two latent spaces some common semantic structures can be constructed. For example, these two latent spaces can be assumed to represent the same concept spaces. That is to say, if the transformation matrices W and V are seen as the weighting coefficients of a set of linear classifiers, and the obtained $\boldsymbol{u}_i$ and $\boldsymbol{z}_j$ are the concept vectors with each entry indicating the membership of one latent concept. Meanwhile, these transformed samples in these two concept spaces interact with each other according to the relation matrix A. A large a_{ij} means the semantic meanings of $\boldsymbol{u}_i$ and $\boldsymbol{z}_j$ have a stronger correlation which could help to infer one latent sample $\boldsymbol{u}_i$ from the other one $\boldsymbol{z}_j$. In other words, it is assumed that the transformed samples in these two latent spaces can be reconstructed by each other based on these underlying correlations revealed by A (see Figure 3 (b) and (c)). Formally, we can use the samples in Z to reconstruct the samples in U based on the elements in A as (see Figure 3 (b))

$$\boldsymbol{u}_i = \frac{1}{\sum_{j=1}^{m} a_{ij}} \sum\nolimits_{j=1}^{m} a_{ij} \boldsymbol{z}_j \tag{1}$$

and similarly, the samples in Z can also be reconstructed as (see Figure 3 (c))

$$\boldsymbol{z}_j = \frac{1}{\sum_{i=1}^{n} a_{ij}} \sum\nolimits_{i=1}^{n} a_{ij} \boldsymbol{u}_i \tag{2}$$

The terms in front of the above two summation operators serve as normalization factors. Eqn. (1) and (2) can be rewritten in more compact matrix formulations as

$$U = Z \cdot A_R^T \tag{3}$$

and

$$Z = U \cdot A_C \tag{4}$$

where A_R and A_C are the row-normalized and column-normalized relation matrices with $[A_R]_{ij} = \frac{a_{ij}}{\sum_{j=1}^{m} a_{ij}}$ and $[A_C]_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}$.

Substitute Eqn. (4) into (3), we can obtain

$$U = U A_C A_R^T \tag{5}$$

Meanwhile according to the linear transformations mentioned above, U can also be represented as

$$U = WX \tag{6}$$

Thus combining Eqn. (5) and (6), we have

$$WX = WXA_C A_R^T \tag{7}$$

The above equation is over-determined and usually there is no exact solutions to W. However, we can use least-square method to solve W.

The least-square solution to W according to Eqn. (7) can be written as

$$\begin{aligned} & W^\star = \arg\min_W ||WX - WXA_C A_R^T||_F^2 \\ & = \arg\min_W tr \left(I - A_C A_R^T\right)^T X^T W^T W X \left(I - A_R A_C^T\right) \end{aligned} \tag{8}$$

Note that our ultimate goal is to learn the distance function between two samples $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ rather than the transformation matrix W, so we have the following squared distance function

$$\begin{aligned} & dist\left(\boldsymbol{x}, \tilde{\boldsymbol{x}}\right) = \left(W\boldsymbol{x} - W\tilde{\boldsymbol{x}}\right)^T \left(W\boldsymbol{x} - W\tilde{\boldsymbol{x}}\right) \\ & = \left(\boldsymbol{x} - \tilde{\boldsymbol{x}}\right)^T W^T W \left(\boldsymbol{x} - \tilde{\boldsymbol{x}}\right) = \left(\boldsymbol{x} - \tilde{\boldsymbol{x}}\right)^T M_X \left(\boldsymbol{x} - \tilde{\boldsymbol{x}}\right) \end{aligned} \tag{9}$$

where $M_X = W^T W \succ 0$ is a symmetric positive-definite matrix. So the objective function of Eqn. (8) is changed to directly solve M_X

$$\begin{aligned} & M_X^\star = \arg\min_{M_X} tr \left(I - A_C A_R^T\right)^T X^T M_X X \left(I - A_R A_C^T\right) \\ & s.t. M_X \succ 0 \end{aligned} \tag{10}$$

This formulation attempts to compute M_X based on the relation matrix A. As discussed in Section 2.1, such a relation comes from users' tagging on the image-sharing website. Unfortunately, these user tags can be quite noisy. Therefore, we would like to introduce a regularization principle to prevent the above formulation from overfitting the noisy tags in the real-world problem.

The motivation of this new principle is to use the image similarity as a regularizer. That is to say if two samples are similar, they ought to have a smaller distance. This motivation is quite intuitive because two similar images tend to be semantically correlated. With this motivation, we can find when M_X in Eqn. (8) becomes an identity matrix I, the distance function reduces to a Euclidean distance between two feature vectors which are extracted from images. If we see the identity matrix I seen as a prior, the obtained M_X ought to be as "close" as possible to it according to the above regularization principle [5].

The "closeness" between M_X and I can be quantified by Bregman divergence between these two symmetric positive-definite matrices [2]. Let $G : \mathcal{S} \to \mathcal{R}$ be a continually-differentiable real-valued and strictly convex function defined on a closed convex set $\mathcal{S}$. Then the Bregman divergence associated with G for $X, Y \in \mathcal{S}$ is

$$D_G\left(X||Y\right) = G(X) - G(Y) - \langle \nabla G(Y), X - Y \rangle \tag{11}$$

In this paper, we use the logdet function $G(X) = -\log det(X)$ over the cone of positive-definite matrices and the corresponding Bregman divergence becomes

$$D_G\left(X||Y\right) = \log \frac{\det\left(Y\right)}{\det\left(X\right)} + tr\left(Y^{-1}X\right) - n \tag{12}$$

By using the above Bregman divergence as the regularizer, the objective function of Eqn. (8) can be rewritten as

$$\begin{aligned} & M_X^\star = \arg\min_{M_X} F(M_X) \\ & = \arg\min_{M_X} tr \left(I - A_C A_R^T\right)^T X^T M_X X \left(I - A_C A_R^T\right) \\ & + \lambda D_G\left(M_X||I\right) \\ & s.t. M_X \succ 0 \end{aligned} \tag{13}$$

where λ is a trading-off parameter. It is not difficult to find that the Bregman divergence is convex in the first argument. Considering the first term in the above objective function is linear in M_X, this objective function is convex and thus has a global optimum solution. We will prove that there exists a closed-form solution to Eqn. (13) here.

Take the derivative of $F(M_X)$ to M_X and set the result to zero, we have

$$\begin{aligned} & \frac{\partial F_X\left(M_X\right)}{\partial M_X} \\ & = \frac{\partial tr \left(I - A_C A_R^T\right)^T X^T M_X X \left(I - A_C A_R^T\right)}{\partial M_X} \\ & + \lambda \frac{\partial}{\partial M_X}\left(-\log\det\left(M_X\right) + tr\left(M_X\right) - d_X\right) \\ & = X\left(I - A_C A_R^T\right)\left(I - A_C A_R^T\right)^T X^T + \lambda\left(-M_X^{-1} + I\right) \\ & = 0 \end{aligned} \tag{14}$$

In the above derivation, we utilize two formulae $\frac{\partial \log\det\left(M_X\right)}{\partial M_X} = M_X^{-1}$ and $\frac{\partial tr\left(M_X\right)}{\partial M_X} = I$.

From Eqn. (14), the minimum point of the objective function (13) can be derived as

$$M_X = \left(I + \lambda^{-1} X \left(I - A_C A_R^T\right) \left(I - A_C A_R^T\right)^T X^T\right)^{-1} \tag{15}$$

It is not difficult to verify this obtained matrix also satisfies the positive-definite constraint in Eqn. (13) since $X\left(I - A_C A_R^T\right) \cdot \left(I - A_C A_R^T\right)^T X^T$ is nonnegative-definite.

Now substitute the above result into Eqn. (9), the distance between two samples $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in R^{d_X}$ is

$$\begin{aligned} &dist\left(\boldsymbol{x}, \tilde{\boldsymbol{x}}\right) \\ &= (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \left(I + \lambda^{-1} X \left(I - A_C A_R^T\right) \left(I - A_C A_R^T\right)^T X^T\right)^{-1} \\ &\cdot (\boldsymbol{x} - \tilde{\boldsymbol{x}}) \end{aligned} \tag{16}$$

In the same manner, we can derive the distance between two samples in $\mathcal{R}^{d_Y}$ as

$$\begin{aligned} &dist\left(\boldsymbol{y}, \tilde{\boldsymbol{y}}\right) \\ &= (\boldsymbol{y} - \tilde{\boldsymbol{y}})^T \left(I + \lambda^{-1} Y \left(I - A_R^T A_C\right) \left(I - A_R^T A_C\right)^T Y^T\right)^{-1} \\ &\cdot (\boldsymbol{y} - \tilde{\boldsymbol{y}}) \end{aligned} \tag{17}$$

2.3 Kernelization

The obtained distance functions (16) and (17) are still linear w.r.t. their input feature spaces. In this section, we will extend them by kernelizing them. We believe such a nonlinearity property of kernelization form can significantly improve the performance of these distance functions due to the underlying nonlinear structure of images just like many successful kernel methods [15].

First a transformation ϕ maps the samples in the input space into a target space in which kernel function $k(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ gives the inner product. Then we can rewrite Eqn. (16) by computing the distance in the target space as

$$\begin{aligned} &dist\left(\phi_X(\boldsymbol{x}), \phi_X(\tilde{\boldsymbol{x}})\right) = (\phi_X(\boldsymbol{x}) - \phi_X(\tilde{\boldsymbol{x}}))^T \\ &\cdot \left(I + \lambda^{-1} \phi_X(X) \left(I - A_C A_R^T\right) \left(I - A_C A_R^T\right)^T \phi_X(X)^T\right)^{-1} \\ &\cdot (\phi_X(\boldsymbol{x}) - \phi_X(\tilde{\boldsymbol{x}})) \end{aligned} \tag{18}$$

where $\phi_X(X) = [\phi_X(\boldsymbol{x}_1), \phi_X(\boldsymbol{x}_2), \cdots \phi_X(\boldsymbol{x}_n)]$.

Applying the Woodbury formula $(I + A \cdot B)^{-1} = I - A(I + B \cdot A)^{-1} B$ into Eqn. (18), we have

$$\begin{aligned} &dist\left(\phi_X(\boldsymbol{x}), \phi_X(\tilde{\boldsymbol{x}})\right) \\ &= (\phi_X(\boldsymbol{x}) - \phi_X(\tilde{\boldsymbol{x}}))^T \cdot \{I - \phi_X(X) \left(I - A_C A_R^T\right) \\ &\cdot \left(\lambda I + \left(I - A_C A_R^T\right)^T \phi_X(X)^T \phi_X(X) \left(I - A_C A_R^T\right)\right)^{-1} \\ &\cdot \left(I - A_C A_R^T\right)^T \phi_X(X)^T\} \cdot (\phi_X(\boldsymbol{x}) - \phi_X(\tilde{\boldsymbol{x}})) \\ &= (\phi_X(\boldsymbol{x}) - \phi_X(\tilde{\boldsymbol{x}}))^T (\phi_X(\boldsymbol{x}) - \phi_X(\tilde{\boldsymbol{x}})) \\ &- (\phi_X(\boldsymbol{x}) - \phi_X(\tilde{\boldsymbol{x}}))^T \phi_X(X) \left(I - A_C A_R^T\right) \\ &\cdot \left(\lambda I + \left(I - A_C A_R^T\right)^T \phi_X(X)^T \phi_X(X) \left(I - A_C A_R^T\right)\right)^{-1} \\ &\cdot \left(I - A_C A_R^T\right)^T \phi_X(X)^T (\phi_X(\boldsymbol{x}) - \phi_X(\tilde{\boldsymbol{x}})) \end{aligned} \tag{19}$$

Incorporating the inner product function $k(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ in the target space into the above equation, we have

$$\begin{aligned} &dist\left(\phi_X(\boldsymbol{x}), \phi_X(\tilde{\boldsymbol{x}})\right) \\ &= k_X(\boldsymbol{x}, \boldsymbol{x}) - 2k_X(\boldsymbol{x}, \tilde{\boldsymbol{x}}) + k_X(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) \\ &- (k_X(\boldsymbol{x}) - k_X(\tilde{\boldsymbol{x}}))^T \left(I - A_C A_R^T\right) \\ &\cdot \left(\lambda I + \left(I - A_C A_R^T\right)^T K_X \left(I - A_C A_R^T\right)\right)^{-1} \\ &\cdot \left(I - A_C A_R^T\right)^T (k_X(\boldsymbol{x}) - k_X(\tilde{\boldsymbol{x}})) \end{aligned} \tag{20}$$

where $k_X(\boldsymbol{x}) = [k_X(\boldsymbol{x}, \boldsymbol{x}_1), k_X(\boldsymbol{x}, \boldsymbol{x}_2), \cdots, k_X(\boldsymbol{x}, \boldsymbol{x}_n)]^T$ is a $n \times 1$ vector which can be seen as a new representation of sample $\boldsymbol{x}$, and $K_X = [k_X(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n$ is the kernel matrix.

Following the same derivation, we also have

$$\begin{aligned} &dist\left(\phi_Y(\boldsymbol{y}), \phi_Y(\tilde{\boldsymbol{y}})\right) \\ &= k_Y(\boldsymbol{y}, \boldsymbol{y}) - 2k_Y(\boldsymbol{y}, \tilde{\boldsymbol{y}}) + k_Y(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{y}}) \\ &- (k_Y(\boldsymbol{y}) - k_Y(\tilde{\boldsymbol{y}}))^T \left(I - A_R^T A_C\right) \\ &\cdot \left(\lambda I + \left(I - A_R^T A_C\right)^T K_Y \left(I - A_R^T A_C\right)\right)^{-1} \\ &\cdot \left(I - A_R^T A_C\right)^T (k_Y(\boldsymbol{y}) - k_Y(\tilde{\boldsymbol{y}})) \end{aligned} \tag{21}$$

where $k_Y(\boldsymbol{y}) = [k_Y(\boldsymbol{y}, \boldsymbol{y}_1), k_Y(\boldsymbol{y}, \boldsymbol{y}_2), \cdots, k_Y(\boldsymbol{y}, \boldsymbol{y}_n)]^T$ and $K_Y = [k_Y(\boldsymbol{y}_i, \boldsymbol{y}_j)]_{i,j=1}^m$ are the counterparts of $K_X(\boldsymbol{x})$ and K_X in the other input space $\mathcal{R}^{d_Y}$.

3. CONNECTION WITH RCA-DISTANCE

In this section, we will reveal the connection between the above proposed multi-label distance metric and the well-known RCA-based distance metric [1]. We will show that the RCA-based distance is only a special case of the proposed distance. This section can be skipped since no problem will be caused to understand the proposed algorithm itself.

Before we reveal this connection, we give a brief review of the RCA-based distance learning. The basic idea is to learn a distance metric from the similarity information provided in the form of chunklets. The samples in the same chunklet belong to the same class. Formally, we are given L chunklets $X = [\boldsymbol{x}_{1,1}, \boldsymbol{x}_{1,2}, \cdots \boldsymbol{x}_{1,n_1}, \cdots, \boldsymbol{x}_{l,1}, \boldsymbol{x}_{l,2}, \cdots \boldsymbol{x}_{l,n_l}, \cdots, \boldsymbol{x}_{L,1}, \boldsymbol{x}_{L,2}, \cdots \boldsymbol{x}_{L,n_L}]$ with chunklet l containing n_l samples. Then, the covariance matrix of the centered patterns in all the chunklets can be computed as

$$C = \frac{1}{n} \sum_{l=1}^{L} \sum_{i=1}^{n_c} (\boldsymbol{x}_{l,i} - m_l)(\boldsymbol{x}_{l,i} - m_l)^T \tag{22}$$

where m_l is the mean of chunklet l and $n = \sum\limits_{l=1}^{L} n_l$ is the total number of samples. With this covariance matrix, the distance function between two samples can be computed as

$$dist_{RCA}\left(\boldsymbol{x}, \tilde{\boldsymbol{x}}\right) = (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T C^{-1} (\boldsymbol{x} - \tilde{\boldsymbol{x}}) \tag{23}$$

To reveal the connection between RCA distance and the proposed multi-label distance, the above covariance matrix C can be transformed with some matrix operation [18] as

$$\begin{aligned} C &= \sum_{l=1}^{L} \sum_{i=1}^{n_l} (\boldsymbol{x}_{l,i} - m_l)(\boldsymbol{x}_{l,i} - m_l)^T \\ &= \sum_{l=1}^{L} X \left(D_l - \tfrac{1}{n_l} 1_l \cdot 1_l^T\right) X^T \\ &= X \left(I - \sum_{l=1}^{L} \tfrac{1}{n_l} 1_l \cdot 1_l^T\right) X^T \end{aligned} \tag{24}$$

where 1_l is the n-dimensional vector with

$$[1_l]_i = \begin{cases} 1, \text{sample} i \in \text{chunklet} l \\ 0, \text{otherwise} \end{cases}$$

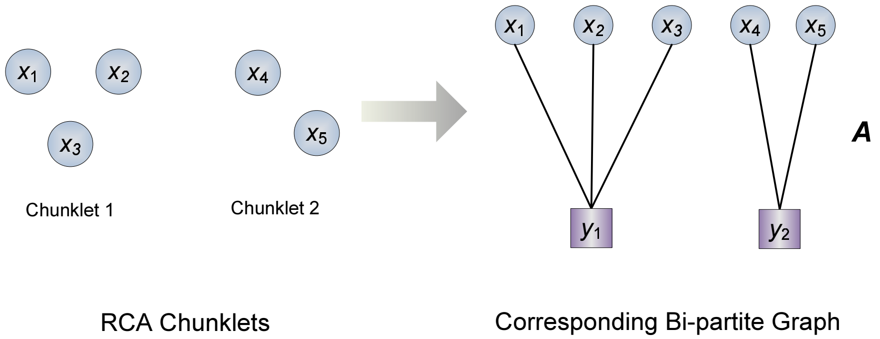


Figure 4: Connection between RCA-based distance learning and the proposed multi-label distance metric learning: the image nodes in the same RCA chunklet are associated with one chunklet node in the representation of bipartite graph. With such a conversion from RCA chunklets to the corresponding bipartite graph, we prove in Section 3 the RCA-based metric distance is only a special case of the proposed multi-label distance metric.

where $D_l = diag\left(1_l\right)$ is an $n \times n$ diagonal matrix with 1_l as its diagonal elements.

The chunklets in the RCA formulation can also be represented by the relation matrix $A = (a_{ij})$ in the bipartite graph like Figure 2, in which $a_{ij} = 1$ if and only if the sample $\boldsymbol{x}_i$ belongs to chunklet j or otherwise $a_{ij} = 0$. Figure 4 illustrates an example of such a conversion from RCA chunklets to the bipartite graph representation. Note that although we can easily convert RCA chunklets into bipartite graph, the reverse conversion does not exist directly. In the following, we will see the RCA distance is only a special case of the proposed distance metric in which each image node can only be associated with one chunklet node in the corresponding bipartite graph.

By substituting the above relation matrix A into the following equation, we can verify that

$$\begin{aligned}
& X\left(I - A_C A_R^T\right)\left(I - A_C A_R^T\right)^T X^T \\
&= X\left(I - \sum_{l=1}^{L} \frac{1}{n_l} 1_l \cdot 1_l^T\right)\left(I - \sum_{l=1}^{L} \frac{1}{n_l} 1_l \cdot 1_l^T\right)^T X^T \\
&= X\left(I - \sum_{l=1}^{L} \frac{2}{n_l} 1_l \cdot 1_l^T + \sum_{l=1}^{L} \sum_{l'=1}^{L} \frac{1}{n_l} \frac{1}{n_{l'}} 1_l \cdot 1_l^T \cdot 1_{l'} \cdot 1_{l'}^T\right) X^T \\
&= X\left(I - \sum_{l=1}^{L} \frac{2}{n_l} 1_l \cdot 1_l^T + \sum_{l=1}^{L} \frac{1}{n_l} 1_l \cdot 1_l^T\right) X^T \\
&= X\left(I - \sum_{l=1}^{L} \frac{1}{n_l} 1_l \cdot 1_l^T\right) X^T
\end{aligned} \tag{25}$$

where $A_R = \left[\begin{array}{cccc} 1_1 & 1_2 & \cdots & 1_L \end{array}\right]$, $A_C = \left[\frac{1}{n_1} 1_1 \frac{1}{n_2} 1_2 \cdots \frac{1}{n_L} 1_L\right]$. In the above derivation, we utilize

$$1_l^T \cdot 1_{l'} = \begin{cases} n_l, \text{when} l = l' \\ 0, \text{otherwise} \end{cases}$$

Accordingly, the distance function in Eqn. (16) can be rewritten as

$$dist\left(\boldsymbol{x}, \tilde{\boldsymbol{x}}\right) = \left(\boldsymbol{x} - \tilde{\boldsymbol{x}}\right)^T \left(I + \lambda^{-1} C\right)^{-1} \left(\boldsymbol{x} - \tilde{\boldsymbol{x}}\right) \tag{26}$$

Compare Eqn. (23) and (26), we can find that the RCA distance is only a special case of the proposed distance metric in a special bi-partite graph in which each image node can only be associated with one chunklet node (see Figure 4 for an example)čň with an exception of an extra term I in $\left(I + \lambda^{-1} C\right)^{-1}$ that serves as a regularizer to respect the prior of Euclidean distance. In the other words, each image sample belongs to one and only one chunklet in RCA distance learning.

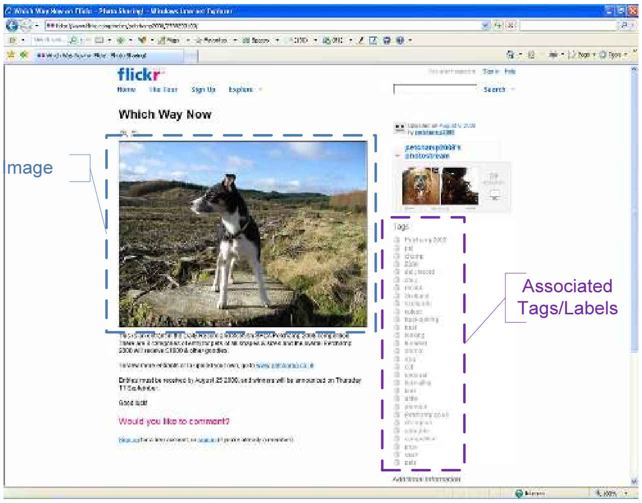
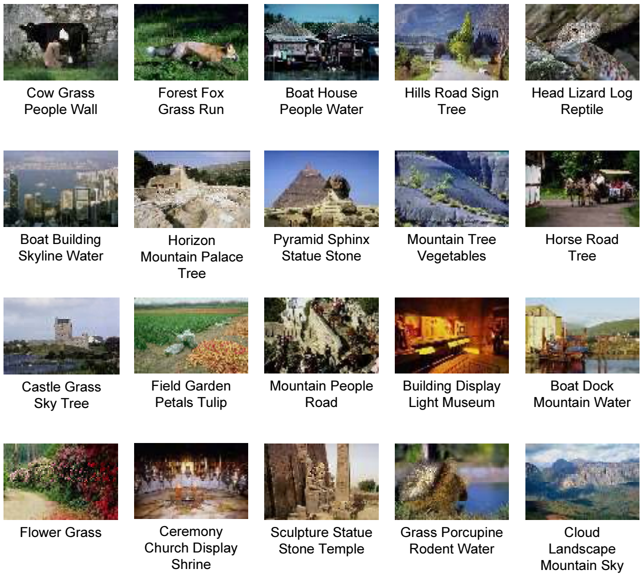


Figure 5: An example image on Flickr website together with its annotated tags/labels.

Unlike the RCA distance learning, the proposed multi-label distance metric removes such a restriction and each image can be associated with more than one labels. Such a multi-label distance metric can be applied to many real-world applications. For example, in an image-sharing website, each image is usually associated with more than one tags labeled by users. By using the proposed algorithm, the learned multi-label distance metric can be directly applied to rank the relevance of the images in a corpus according to their distances to the query.

4. EXPERIMENTS

This section empirically evaluates the proposed multi-label distance metric to retrieve the images compared with other representative distance metrics.



Cow Grass People Wall

Forest Fox Grass Run

Boat House People Water

Hills Road Sign Tree

Head Lizard Log Reptile

Boat Building Skyline Water

Horizon Mountain Palace Tree

Pyramid Sphinx Statue Stone

Mountain Tree Vegetables

Horse Road Tree

Castle Grass Sky Tree

Field Garden Petals Tulip

Mountain People Road

Building Display Light Museum

Boat Dock Mountain Water

Flower Grass

Ceremony Church Display Shrine

Sculpture Statue Stone Temple

Grass Porcupine Rodent Water

Cloud Landscape Mountain Sky

Figure 6: Some example images in the experiments together with their associated user tags.

4.1 Dataset

We evaluate the proposed distance learning algorithms on two different datasets for image retrieval.

Corel Image Dataset - The first one is the benchmark Corel image dataset. It contains 5,460 images which are annotated with 393 different tags provided by a group of professional users. Each image is annotated by average 2.88 tags and each tag is assigned to average 39.98 images. The dataset is split into two parts - 3,000 images as the training set to learn the distance and the remaining 2,460 images as the queries to test the retrieval performance. Some example images are shown in Figure 6 together with annotated user tags.

Flickr Dataset - The second one is a realistic dataset crawled from an image-sharing website Flickr. It includes 32,575 images together with 4,100 tags on these images provided by users (See Figure 5 for an example). In this set, each image has average 10.51 tags and each tag is assigned to 92.05 images on average. For performance evaluation, we also split the set into two parts. 15,000 images are used as training set and the remaining 17,575 images are used to simulate user queries.

To represent the visual content of images, we extract 225-dimensional color moment features for each image. In detail, the image is divided into 5 × 5 regular grid and the first three orders of color moments in each grid are extracted from each component of RGB color space. On the other hand, the tag feature vectors are represented by 1/0 vectors with the elements being 1 for the annotated tags and the other element being 0 otherwise.

4.2 Evaluation Measure

Since we focus on image retrieval by the learned distance metrics, a ranking-based evaluation measure, Normalized Discounted Cumulative Gain at top k ($NDCG@k$) [9], is adopted to evaluate the retrieval performance. In contrast to other measures such as precision and recall that only measure the accuracies of retrieved results, $NDCG@k$ measures the different levels of relevance and prefers the retrieved ranking results that follow the actual relevance order. Thus this evaluation measure can better reflect the users' requirement of ranking the most relevant images at top in a real retrieval system. The formula of $NDCG@k$ can be computed as

$$NDCG@k = \frac{1}{Z}\sum_{p=1}^{k}\frac{2^{s(p)}-1}{\log(1+p)} \tag{27}$$

where $s(p)$ is the function that represents reward given to the retrieved image at position p, Z is a normalization term derived from the perfect ranking of top k images so that it can normalize $NDCG@k$ to be $[0,1]$. The above summation is computed from position 1 to k in the ranked list according to the distance metric in the experiments.

The reward function $s(p)$ is measured by the relevance between the image at position p and the proposed query. We use the cosine similarity between the tag vectors of the image at position p of retrieved list and the query image as their relevance measure

$$s(p) = \frac{\langle v_p, v_q\rangle}{||v_p||_2||v_q||_2} \tag{28}$$

v_p and v_q are the tag vectors of the pth image in the ranking list and the query image, respectively. Each entry in these tag vectors indicates if the corresponding tag occurs in the image. $\langle\cdot,\cdot\rangle$ is their inner product, and $||\cdot||_2$ is the 2-order vector norm.

For each image used as user query, we compute its $NDCG@k$ and then report the average over the whole testing set as the performance measure of the learned distance metrics.

4.3 Compared Algorithms

We compare the following distance metrics to evaluate the proposed metric.

Euclidean - It directly computes the Euclidean distance between visual features of the images. This distance metric only utilizes the visual information and no textual information such as user tags is considered.

RCA (Relevant Component Analysis) - As discussed in Section 3, RCA learns the distance metric from chunklets each of which clusters a set of similar images together. To provide these chunklets to learn the RCA distance metric, we adopt a classic co-clustering algorithm [6] to cluster the bipartite graph like Figure 2 to obtain some image chunklets. Then the RCA distance learning algorithm is applied with these chunklets to learn the distance metric for image retrieval.

NCA (Neighborhood Components Analysis) - It learns a Mahalanobis distance metric by maximizing a stochastic variant of the leave-one-out KNN score on the training set. For details, please refer to [8].

ML-DML (Multi-Label Distance Metric Learning) - It is the proposed multi-label distance metric learning in Section 2.2.

KML-DML (Kernel ML-DML) - It is the kernel extension of ML-DML in Section 2.3.

Next we will compare the above distance learning algorithms on Corel and Flickr datasets. Note that although there exist other advanced distance learning algorithms, such as ITML [5] and LMNN [21], it is intractable to apply them to train distance model on more than hundreds of images

Table 1: Comparison of $NDCG@k$ measure for Euclidean, RCA, NCA, ML-DML and KML-DML algorithms on Corel Dataset, where k is from 300 to 3000.

$NDCG@k$	Euclidean	RCA	NCA	ML-DML	KML-DML
300	0.1302	0.1369	0.1582	0.1990	**0.2112**
600	0.1733	0.1808	0.2019	0.2534	**0.2698**
900	0.2221	0.2287	0.2512	0.3072	**0.3253**
1200	0.2773	0.2850	0.3067	0.3649	**0.3851**
1500	0.3357	0.3422	0.3618	0.4233	**0.4445**
1800	0.3930	0.3990	0.4161	0.4787	**0.4990**
2100	0.4480	0.4535	0.4687	0.5317	**0.5514**
2400	0.5028	0.5069	0.5209	0.5821	**0.6001**
2700	0.5531	0.5568	0.5687	0.6302	**0.6479**
3000	0.6059	0.6088	0.6177	0.6796	**0.6965**

Table 2: Comparison of $NDCG@k$ measure for Euclidean, RCA, NCA, ML-DML and KML-DML algorithms on Flickr Dataset, where k is from 1500 to 15000.

$NDCG@k$	Euclidean	RCA	NCA	ML-DML	KML-DML
1500	0.2644	0.2725	0.2888	0.3360	**0.3695**
3000	0.3251	0.3279	0.3387	0.3843	**0.4199**
4500	0.3799	0.3812	0.3868	0.4301	**0.4677**
6000	0.4353	0.4348	0.4360	0.4773	**0.5170**
7500	0.4913	0.4897	0.4872	0.5261	**0.5679**
9000	0.5492	0.5463	0.5414	0.5790	**0.6233**
10500	0.6097	0.6066	0.6028	0.6398	**0.6870**
12000	0.6696	0.6685	0.6686	0.7063	**0.7568**
13500	0.7296	0.7304	0.7338	0.7740	**0.8281**
15000	0.7907	0.7923	0.7953	0.8364	**0.8941**

due to their high computational complexity, so we will not compare with them in this paper.

4.4 Performance Comparison

Experiment results conducted on the Corel and Flickr datasets are shown in Table 1 and Table 2. We compare the proposed ML-DML and KML-DML algorithms with Euclidean, RCA and NCA algorithms in terms of $NDCG@k$, where k varies on these two different datasets. We can find the Euclidean distance has the worst performance since it does not utilize any tagging information provided by users. The other four algorithms have gained better performance with the help of these user tags. Among all the compared distance metrics, the proposed ML-DML and KML-DML emerges to gain the best performance since they combine the information from both visual content of images and the associated user tags in an integrated manner mentioned above. It is quite different from RCA and NCA in which a preprocessing step is required to construct “chunklets” from the user tags.

We illustrate some retrieval results by different distance metrics in Figure 7. These images are ranked based on their distance to the query image. The images with smaller distances are ranked higher. These examples also illustrate the better performance of KML-DML and ML-DML compared to other distance learning algorithms.

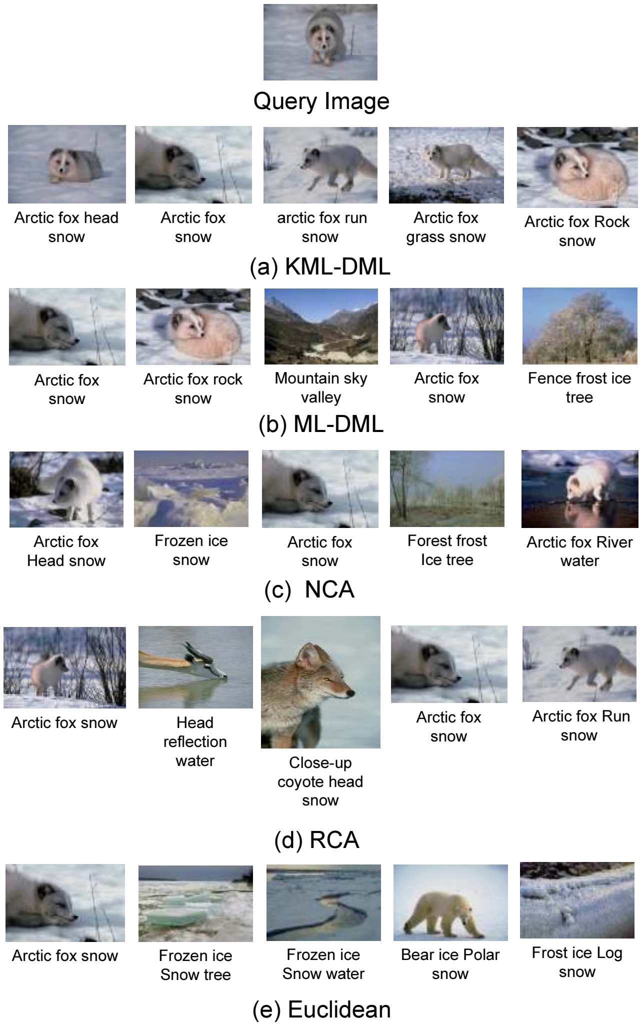
4.5 Application to Image Annotation

In addition to image retrieval, the distance metrics can be applied to automatic image annotation. Given an image without tags, those tags associated with the k nearest images can be assigned as the tags of this image. For example, Figure 8 illustrates the tagging annotation of an example image with different k from 1 to 5 by different distance metrics. We can see KML-DML and ML-DML perform best among these distances on this example of image annotation.

In Figure 9 and 10, we also quantitively compare the annotation performances by these distance metrics in terms of relative improvements of precision and recall compared to Euclidean metric on the Flickr dataset. We can find the proposed KML-DML and ML-DML outperform NCA and RCA whatever k is used.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel multi-label distance metric learning algorithm which integrates the visual content of images and their associated user tags in the image-sharing website. It directly mines the semantic information provided by user tags together with the assumption that the visual similar images ought to have a smaller semantic distance. The visual information also serves to reduce the overfitting risk due to the noise tagging information via a Bregman regularizer. We formulate to learn distance by an optimization problem which can be efficiently solved by a closed-form solution. We also enhance the proposed multi-label distance into kernelization form so that nonlinearity is incorporated. Experimental evaluation on the Corel and Flickr datasets shows the most competitive performance of the proposed distance metrics compared to Euclidean, RCA and NCA distance learning. In addition, we also apply the learned distance metrics to automatic image annotation. ML-DML and KML-DML outperform other metrics again.



(a) KML-DML

(b) ML-DML

(c) NCA

(d) RCA

(e) Euclidean

Figure 7: Top 5 retrieved images by KML-DML, ML-DML, NCA, RCA and Euclidean. For each distance metric, the images are ranked by their corresponding distance to the query image. The image with smaller distance is ranked higher.

The proposed distance learning algorithm aims at computing the image distances, however, as indicated in Eqn. (16) and Eqn. (17), this algorithm can simultaneously learn a couple of distance metrics on two spaces (i.e., the image space and the tag space). Therefore, it unifies the image distance learning and the associated tag distance learning (like Google Distance [4] in an integrated framework. But due to the limited space in this paper, we delay the discussion and empirical evaluation on tag distance in our future work.

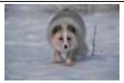
Image to be annotated					
k	1	2	3	4	5
KML-DML	arctic fox head snow	arctic fox head snow	arctic fox head snow run	arctic fox head snow run grass	arctic fox head snow run grass rock
ML-DML	arctic fox snow	arctic fox snow rock	arctic fox snow rock mountain sky valley	arctic fox snow rock mountain sky valley	arctic fox snow rock mountain sky valley fence frost ice tree
NCA	arctic fox head snow	arctic fox head snow frozen ice	arctic fox head snow frozen ice	arctic fox head snow frozen ice forest frost tree	arctic fox head snow frozen ice forest frost tree river water
RCA	arctic fox snow	arctic fox snow head reflection water	arctic fox snow head reflection water close-up coyote	arctic fox snow head reflection water close-up coyote	arctic fox snow head reflection water close-up coyote run
Euclidean	arctic fox snow	arctic fox snow frozen ice tree	arctic fox snow frozen ice tree water	arctic fox snow frozen ice tree water bear polar	arctic fox snow frozen ice tree water bear polar frost log

Figure 8: Image annotation by the k nearest images. The tags in the nearest k (k from 1 to 5) images are assigned to the image to be annotated. Tags with purple color are relevant to the target image.

6. REFERENCES

[1] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *Proc. of International Conference on Machine Learning*, 2003.

[2] L. Bregman. The relaxation method of find the common point of convex sets and its application to the solution of problems in convex programming. In *USSR Comp. Mathematics and Mathematical Physics*, 1967.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM International Conference on Image and Video Retrieval*, 2009.

[4] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370 – 383, March 2007.

[5] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proc. of International Conference on Machine Learning*, 2007.

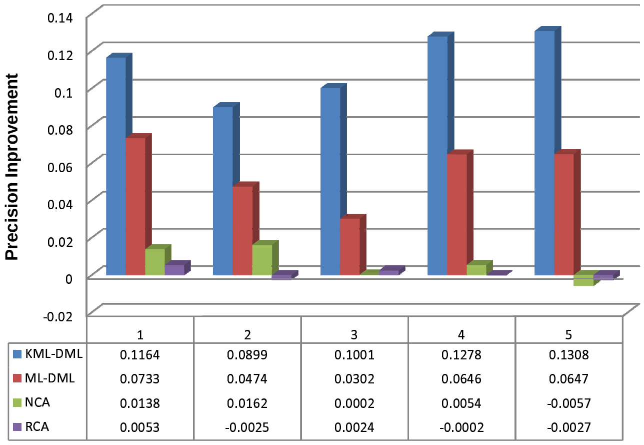
[6] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.

[7] S. C. H. Hoi, W. Liu, and S.-F. Chang. Semi-supervised distance learning for collaborative image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[8] G. H. J. Goldberger, S. Roweis and R. Salakhutdinov. Neighbourhood components analysis. In *Proc. of Advanced Neutral Information Processing System*, 2004.

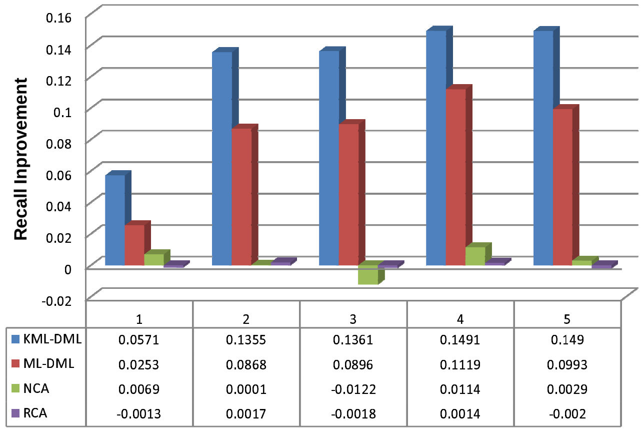
[9] K. Jarvelin and J. Kekalainen. Ir evaluation methods for retrieving highly relevant documents. In *Proc. of International ACM SIGIR Conference on Research and Development on Information Retrieval*, 2000.

[10] B. Long, X. Wu, Z. Zhang, and P. S. Yu. Unsupervised learning on k-partite graphs. In *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.



	1	2	3	4	5
KML-DML	0.1164	0.0899	0.1001	0.1278	0.1308
ML-DML	0.0733	0.0474	0.0302	0.0646	0.0647
NCA	0.0138	0.0162	0.0002	0.0054	-0.0057
RCA	0.0053	-0.0025	0.0024	-0.0002	-0.0027

Figure 9: Relative precision improvement of image annotation by the nearest k (k from 1 to 5) images on Flickr dataset compared to Euclidean distance.



	1	2	3	4	5
KML-DML	0.0571	0.1355	0.1361	0.1491	0.149
ML-DML	0.0253	0.0868	0.0896	0.1119	0.0993
NCA	0.0069	0.0001	-0.0122	0.0114	0.0029
RCA	-0.0013	0.0017	-0.0018	0.0014	-0.002

Figure 10: Relative recall improvement of image annotation by the nearest k images on Flickr dataset compared to Euclidean distance.

[11] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proc. of International ACM Conference on Multimedia*, Augsburg, Germany, September 2007.

[12] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional multi-label active learning with an efficient online adaptation model for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[13] M. Rege, M. Dong, and J. Hua. Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering. In *Proc. of International Conference on World Wide Web*, 2008.

[14] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *Proc. of Advanced Neutral Information Processing System*, 2004.

[15] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Recognition*. Cambridge University Press, 2004.

[16] L. Si, R. Jin, S. C. Hoi, and M. R. Lyu. Collaborative image retrieval via regularized metric learning. *ACM Multimedia Systems Journal*, 2006.

[17] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[18] I. W. Tsang, P.-M. Cheung, and J. T. Kwok. Kernel relevant component analysis for distance metric learning. In *Proc. of International Joint Conference on Neural Networks*, 2005.

[19] C. Wang, L. Zhang, and H.-J. Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*, Singapore, July 2008.

[20] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[21] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Proc. of NIPS*, 2005.

[22] Q. Weinberger, M. Slaney, and R. V. Zwol. Resolving tag ambiguity. In *Proc. of International ACM Conference on Multimedia*, 2008.

[23] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Proc. of Advanced Neutral Information Processing System*, 2003.

[24] J. Yu and Q. Tian. Semantic subspace projection and its application in image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(4):544 – 548, April 2008.