2-2011

Distance Metric Learning from Uncertain Side Information for Automated Photo Tagging

Lei WU
University of Science and Technology of China

Steven C. H. HOI
Singapore Management University, CHHOI@smu.edu.sg

Rong JIN
Michigan State University

Jianke ZHU
Zhejiang University

Nenghai YU
University of Science and Technology of China

DOI: https://doi.org/10.1145/1899412.1899417

Published in ACM Transactions on Intelligent Systems and Technology, Volume 2, Issue 2, February 2011, Article number 13
https://doi.org/10.1145/1899412.1899417

Distance Metric Learning from Uncertain Side Information for Automated Photo Tagging

Lei Wu
University of Science and Technology of China
Steven C.H. Hoi
Nanyang Technological University
Rong Jin
Michigan State University
Jianke Zhu
Zhejiang University
Nenghai Yu
University of Science and Technology of China

Automated photo tagging is an important technique for many intelligent multimedia information systems, e.g. smart photo management system and intelligent digital media library. To attack the challenge, several machine learning techniques have been developed and applied for automated photo tagging. For example, supervised learning techniques have been applied to automated photo tagging by training statistical classifiers from a collection of manually labeled examples. Although the existing approaches work well for small testbeds with relatively small number of annotation words, due to the long-standing challenge of object recognition, they often perform poorly in large-scale problems. Another limitation of the existing approaches is that they require a set of high quality labeled data, which is not only expensive to collect but also time consuming. In this paper, we investigate a social image based annotation scheme by exploiting *implicit* side information that is available for a large number of social photos from the social web sites. The key challenge of our intelligent annotation scheme is how to learn an effective distance metric based on *implicit* side information (visual or textual) of social photos. To this end, we present a novel "Probabilistic Distance Metric Learning" (PDML) framework, which can learn optimized metrics by effectively exploiting the *implicit* side information vastly available on the social web. We apply the proposed technique to photo annotation tasks based on a large social image testbed with over 1 million tagged photos crawled from a social photo sharing portal. Encouraging results show that the proposed technique is effective and promising for social photo based annotation tasks.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Manuscript received Feb 24, 2010; revised April 01, 2010; accepted August 13, 2010.
Contact author's address: Steven C.H. Hoi, School of Computer Engineering, Nanyang Technological University, Singapore 639798, e-mail: chhoi@ntu.edu.sg, Lei Wu, MOE-MS Keynote Lab of MCC, University of Science and Technology of China, P.R. China, e-mail: leiwu@live.com, Rong Jin, Department of Computer Science and Engineering, 3115 Engineering Building, Michigan State University, East Lansing, MI 48824, U.S.A., e-mail: rongjin@cse.msu.edu, Jianke Zhu, Zhejiang University, P.R. China, e-mail: jianke.zhu@gmail.com, Nenghai Yu, MOE-MS Keynote Lab of MCC, University of Science and Technology of China, P.R. China, e-mail: ynh@ustc.edu.cn

General Terms: Algorithms, Experimentation
Additional Key Words and Phrases: Automated photo tagging, distance metric learning, uncertain side information, social images, content-based image retrieval

1. INTRODUCTION

Although content-based image retrieval has been extensively studied [Smeulders et al. 2000; Lew et al. 2006], searching image and photo by textual queries remains one of the most common and imperative functions for most intelligent multimedia systems. For many real-world multimedia systems, raw images and photos are often not associated with text labels or human tags. Automated image annotation thus becomes an important technique to make massive collections of unlabeled images and photos searchable by existing text indexing and retrieval solutions.

In general, an image annotation task is to assign a set of text labels or semantic tags to a novel image based on its visual (and textual if any) content. A typical image annotation approach usually requires two key steps. One is to extract visual features to represent the images [Lowe 2004], and the other is to build accurate classification models from the training images and employ them to predict tags/lables for the query/test images [Carneiro et al. 2006]. Over the past decade, significant efforts have been expended for automated image annotation and object recognition tasks in several areas, including multimedia, computer vision, image processing, and machine learning [Jeon et al. 2003; Smeulders et al. 2000; Lew et al. 2006].

Despite encouraging progresses, most image annotation methods work well on small-sized dataset with high quality training data, but often fail when it comes to large-scale real-world applications for photo tagging due to the well-known semantic gap between low-level image features and high-level semantic concepts. Besides the challenge arising from the semantic gap, it is also expensive and time-consuming to collect a large set of manually-labeled training data for the conventional methods. Therefore, it is urgent to develop new effective paradigms for automated photo tagging beyond the traditional approaches.

Recently, due to the popularity of social networks and social web, massive tagged images have been available on the web, which are referred to as "*social images/photos*". Unlike typical WWW images [Hoi and Lyu 2004], social images often contain manually-labeled tags and rich user-generated contents, which offer a new opportunity to resolve some long-standing challenges in multimedia, e.g., the semantic gap. In this paper, we investigate an emerging retrieval-based annotation paradigm for automated photo tagging by mining massive social images freely available on the web. The basic idea is to first retrieve a set of most similar images for a test photo from the social image repository, and then assign the test photo with the most popular tags associated with the set of similar social images [Wang et al. 2006].

The crux of a retrieval-based annotation paradigm is to *accurately* find the set of similar images. It mainly relies on two key components: (1) image representation by extracting salient visual features from images, and (2) distance measure for computing the dissimilarity between the two images based on the extracted features. In this paper, we focus on the second challenge by learning an optimal metric for

distance measure, known as "Distance Metric Learning" (DML) [Xing et al. 2002].

Existing DML methods work only with *explicit* side information, which is given either in the forms of class labels [Weinberger et al. 2006; Goldberger et al. 2005] or pairwise constraints [Xing et al. 2002; Bar-Hillel et al. 2005; Hoi et al. 2006]. Besides, existing DML methods also assume that the given side information is clean and perfect. Such assumptions seldom hold in a real application. For example, in our application, the tags and contents generated by users for images are often erroneous, and more importantly cannot be used directly as the explicit side information. This motivates us to study a new approach of distance metric learning from uncertain/implicit side information.

To this end, in this paper, we present a novel *Probabilistic Distance Metric Learning* (PDML) framework, which aims to learn distance metrics from noisy and uncertain side information for automated photo tagging tasks. The proposed framework consists of two steps: (1) an unsupervised learning approach for discovering probabilistic side information from hidden erroneous and implicit side information contained in rich user-generated content of social image data; and (2) a PDML approach for learning an optimal distance metric from probabilistic side information.

In summary, the key contributions of this paper include: (1) a retrieval-based annotation scheme powered by a novel DML technique for automated photo tagging; (2) a novel probabilistic DML framework to learn metrics from *erroneous* and *implicit* side information; (3) two effective PDML algorithms, pRCA and pDCA, to learn optimal metrics from probabilistic side information; (4) extensive experiments to verify the efficacy of our algorithms in comparison to a number of state-of-the-art DML algorithms for automated photo annotation tasks.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 presents an overview of the proposed DML framework for automated photo annotation, and proposes solutions for discovering implicit constraints from social photo repositories. Section 4 proposes the probabilistic DML method and gives two efficient algorithms, i.e., probabilistic Relevance Component Analysis (pRCA) and probabilistic Discriminative Component Analysis (pDCA). Section 5 discusses the application of PDML to automated photo tagging. Section 6 presents the experimental results and Section 7 concludes this work.

2. RELATED WORK

Our work is mainly related to two groups of research. One is the group of studies on exploring web/social photo repositories for image annotation and object recognition [Russell et al. 2008; Torralba et al. 2008; Yan et al. 2008]. The other is related to the group of DML studies [Bar-Hillel et al. 2005; Si et al. 2006]. We briefly review some representative work in both sides.

2.1 Automated Photo Tagging

Automated image/photo annotation has been actively studied over the past decade in multimedia community. Among a variety of conventional approaches, a widely-studied paradigm is the supervised classification approach, in which classification models, such as SVM [Fan et al. 2004], are trained from a collection of human-labeled training data for a set of predefined semantic concept/object categories [Carneiro et al. 2006; Carneiro and Vasconcelos 2005; Duygulu et al. 2002;

Wang et al. 2008]. Besides, semi-supervised learning methods are also explored in recent literature [Li and Sun 2006; He and Zemel 2008].

Recent years have witnessed a surge of emerging interests in exploring web photo repositories for image annotation/object recognition problems. One promising approach is the retrieval-based (or termed "search-based") paradigm [Russell et al. 2008; Wang et al. 2006; Torralba et al. 2008; Wang et al. 2008]. Russell et al. [Russell et al. 2008] built a large collection of web images with ground truth labels for helping object recognition research. Wang et al. [Wang et al. 2006] proposed a fast search-based approach for image annotation by some efficient hashing technique. Torralba et al. [Torralba et al. 2008] proposed efficient image search and scene matching techniques for exploring a large-scale web image repository. These studies are usually focused on techniques for fast indexing and search, while we focus on learning effective distance metrics from erroneous and implicit side information. Yan et al. [Yan et al. 2008] proposed a learning based method for improving the efficiency of manual image annotation with the hybrid of tagging and borrowing. Our work differs from theirs by focusing on fully automated photo annotation. Besides, we also notice there are some related work that also learned distance metrics from tagged media collection, such as [Qi et al. 2009; Wang et al. 2008]. Our study differs from them by emphasizing metric learning from uncertain side information.

2.2 Distance Metric Learning

From a machine learning point of view, our work is closely related to DML studies. Firstly, we review some basics of DML. Given a set of n data examples $X = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ in d-dimensional vector space, the Mahalanobis distance between any two examples x_i and x_j is defined as:

$$d_M(x_i, x_j) = \sqrt{((x_i - x_j)^\top M(x_i - x_j))} \tag{1}$$

where M is a positive semi-definite matrix that satisfies the property of valid metric and can be decomposed as $M = A^\top A$. The goal of DML is to find an optimal Mahalanobis metric M from training data (side information) that can be either class labels or general pairwise constraints [Xing et al. 2002].

In literature, DML studies can be roughly divided into two major categories. One is to learn metrics with explicit class labels, such as Neighbourhood Components Analysis (NCA) [Goldberger et al. 2005], which are often studied for classification [Fukunaga 1990; Globerson and Roweis 2005; Weinberger et al. 2006; Yang et al. 2006]. The other is to learn metrics from pair-wise constraints that are mainly used for clustering and retrieval. Examples include RCA [Bar-Hillel et al. 2005] and Discriminative Component Analysis [Hoi et al. 2006], amongst others [Xing et al. 2002]. Our work is more related to the second category, though some methods in the former category could be converted to the latter.

Lots of research studies focus on learning more effective distance metrics with the assistance of the high level semantic from the side information such as pairwise constraints [Xing et al. 2002; Hoi et al. 2006; Weinberger et al. 2006; Davis et al. 2007; Hoi et al. 2008; Jin et al. 2009]. An earlier and well-known DML approach was proposed by Xing et al. [Xing et al. 2002], who formulated the task as a convex optimization problem. The major drawback of their work is computational ineffi-

ciency for large scale dataset. Later, RCA was proposed [Bar-Hillel et al. 2005] to learn metrics with equivalent/relevant constraints, which is simple and efficiency. Discriminant Component Analysis (DCA) further improves RCA by incorporating negative constraints [Hoi et al. 2006]. Most recently, regularized DML and semi-supervised DML algorithms were also studied [Si et al. 2006; Hoi et al. 2008], which were often formulated as an SDP problem and again difficult to be used in large applications. The existing DML algorithms are restricted to rely on explicit pair-wise constraints. Our probabilistic DML overcomes this limitation by exploiting implicit side information, in particular the user-generated content for images, in a probabilistic learning framework.

2.3 Relevant Component Analysis

Here we review a well-known and effective DML technique, i.e., Relevant Component Analysis (RCA) [Bar-Hillel et al. 2005], since it is highly related to our work. The basic idea of RCA is to identify and down-scale global unwanted variability within the data. In particular, RCA suggests to change the feature space used for data representation by a global linear transformation in which relevant dimensions are assigned with large weights. More formally, given a set of data examples $X = \{x_i\}_{i=1}^n$ and a collection of pairwise constraints indicating whether two data examples are similar (or dissimilar). RCA forms a set of m "**chunklets**" $C_j = \{x_{ji}\}_{i=1}^{n_j}$ where $j = 1, \ldots, m$. Each *chunklet* is defined as a group of data examples linked together by similar pair-wise constraints ("must-link").

The optimal transformation by RCA is then computed as $A = \hat{C}^{-1/2}$ and the Mahalanobis matrix is equal to the inverse of the average covariance matrix of chunklets, i.e., $M = \hat{C}^{-1}$, where $\hat{C}$ is defined as follows:

$$\hat{C} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{m_j} (x_{ji} - \mu_j)(x_{ji} - \mu_j)^T \tag{2}$$

where μ_j denotes the mean of j^{th} chunklet, x_{ji} denotes the i^{th} example in the j^{th} chunklet and n is the total number of examples. RCA is simple, efficient, and easy to implement. Similar to other conventional DML techniques, RCA also requires a set of explicit "positive" pairwise constraints provided for the learning task, which limits its application when the side information is given implicitly.

2.4 Discriminative Component Analysis

Discriminative Component Analysis (DCA) [Hoi et al. 2006] aims to learn from both positive constraints and negative constraints. Here a positive constraint indicates two instances are in the same chunklet, and a negative one indicates two instances are in different chunklets. For each chunklet j, a set of discriminative chunklets is formed if there is at least one negative constraint with the j^{th} chunklet.

DCA learns the optimal transformation A by maximizing the total variance between discriminative chunklets and minimizing the total variance of data instances in the same chunklet simultaneously, which can be formally formulated below:

$$\max_A J(A) = \frac{|A^T \hat{C}_b A|}{|A^T \hat{C}_w A|} \tag{3}$$

$$\hat{C}_b = \frac{1}{n_b}\sum_{j=1}^{n}\sum_{i\in D_j}(m_j - m_i)(m_j - m_i)^T, \quad \hat{C_w} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ji} - m_j)(x_{ji} - m_j)^T$$

where D_j is the discriminative set for the j^{th} chunklet, m_j is the mean vector of the j^{th} chunklet, and n_b is the cardinality of all the discriminative sets.

2.5 Certain Side Information V.S. Uncertain Side Information

Side information is critical to any distance metric learning algorithm. It typically appears in the forms of pairwise constraints, which a positive (negative) constraint indicates whether a pair of samples are similar (or dissimilar). Traditional DML methods assume that perfect side information is provided explicitly, which is referred to as *certain side information*. In most studies, certain side information is cast in the hard pairwise constraints that indicate two examples are either absolutely similar or absolutely dissimilar. Besides, certain side information is usually assumed to be perfect without any error. The manual nature of certain side information makes it expensive to collect. These limitations restrict the application of certain side information.

In our study, we focus on learning a distance metric from *uncertain side information* that allows the *uncertainty* when generating the side information, which differs from the certain side information in several aspects. First, it is often generated automatically, e.g. derived from the user-generated content of social images available on the web. Thus, uncertain side information is often much cheaper to acquire than certain side information. Second, it adopts "soft" pairwise constraints, in which each pairwise constraint is associated with a confidence/uncertainty. It is the soft constraints that allow us to better deal with the potentially noisy constraints.

3. METRIC LEARNING FRAMEWORK FOR AUTOMATED PHOTO TAGGING

3.1 Overview

We first give an overview of the proposed semantic metric learning framework for learning metrics from social image data. Figure 1 shows a flowchart illustrating the proposed framework with application to automated photo tagging.

In the figure, the right column shows a retrieval-based photo tagging solution. Specifically, given a novel photo, the idea of the retrieval-based tagging approach is to firstly perform a similarity search for finding top k most similar photos from the social photo repository, and then annotate the novel photo with top t ranked tags associated with the k retrieved photos. Our main effort focuses on learning an effective metric to reduce semantic gap for the similarity based search process, which is shown in the left panel of the flowchart. Below we discuss the main ideas of our metric learning framework.

Since no explicit side information is available, we cannot directly apply regular DML techniques. Therefore, the first step towards DML is to discover *possible* side information from training data, which is essential to DML. In another words, we wish to find some forms of side information, which could indicate how likely two social image examples are similar or dissimilar. One solution is to discover some "**chunklets**" (similar to RCA) from training data such that images in the same chunklets are similar to each other, and images in different chunklets could

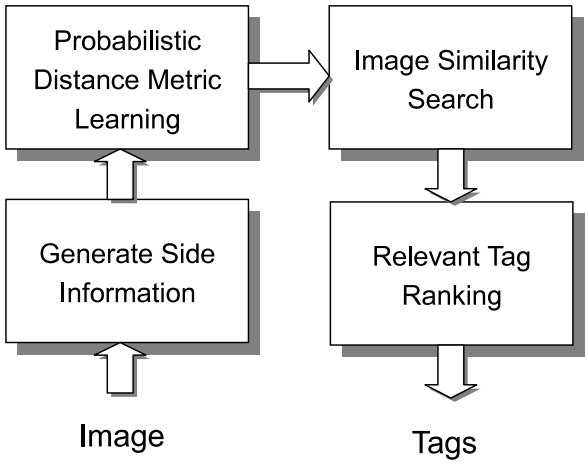


Fig. 1. Flowchart illustrating the proposed metric learning scheme for annotation.

be similar or dissimilar, up to the similarity of the two associated chunklets. Since such chunklets are not explicitly available (also cannot be easily formed as RCA), we refer to them as "*latent chuklets*". Intuitively, a latent chunklet can be viewed as a common semantic topic shared by the social images in the chunklet. Thus, it is possible that one image belongs to multiple chunklets.

To find the latent chunklets effectively and precisely, we propose a graphical model to estimate the probabilities of assigning an image to the latent chunklets. We refer to this step as "Latent Chunklet Estimation" (LCE) step. By LCE, we obtain side information in the form of latent chunklets with probabilistic assignments, which we refer to as "probabilistic side information" or "uncertain side information". Finally, the last step of our semantic metric learning is to find an optimal metric from the probabilistic side information. In this paper, we propose two PDML algorithms, i.e., probabilistic relevant component analysis (pRCA) and probabilistic discriminative component analysis (pDCA), for solving the PDML tasks effectively.

Next we first present the algorithms for latent chunklet estimation followed by the proposed pRCA and pDCA algorithms in the subsequent section.

3.2 Latent Chunklet Estimation for Social Image Modeling

Typically a social image contains rich information, such as tags, title, description, comments, visual content, etc. In this paper, we propose two approaches for discovering side information of latent chunklets from rich contents of social images. One is a graphical model approach, and the other is a clustering based approach. For simplicity, we focus on exploring two key types of information, i.e., textual and visual. It is not difficult to engage additional information in our framework.

3.2.1 *Latent chunklet definition.* First of all, we assume that there are m latent chunklets available, each of them represents a hidden topic z_i, in which both visual images and associated textual metadata (e.g. tags) in the chunklets are generated

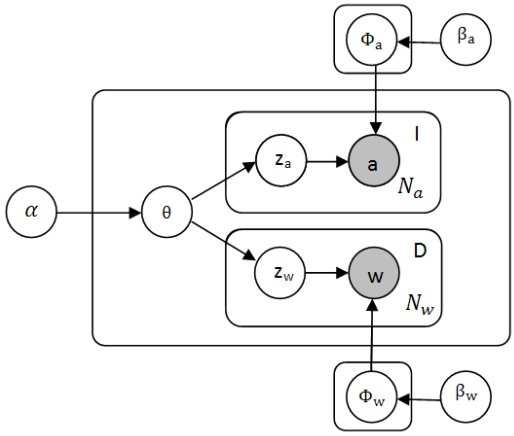


Fig. 2. Graphical model approach for social image modeling

from the hidden topic. Figure 2 shows the graphical model for social image modeling. The upper part of the graph represents the visual model. The images can be represented by some local feature descriptor, e.g. bag of visual words representation [Lowe 2004], and each visual word a is generated from certain topic z_a by a multinomial distribution ϕ_a^z. On the left side, θ is a Dirichlet distribution with hyper parameter α. The lower part of the graph represents the textual model generating textual tags, in which w represents the tags. For simplicity, we also assume that the tags are generated from a multinomial distribution ϕ_w^z parameterized by the topic z_w. Thus, a topic z contains two parts, i.e., $z = [z_a, z_w]$.

Our goal is to estimate the hidden distribution $P(z_a|I)$, the probability of an image I belonging to a certain topic z_a, and the hidden distribution $P(z_w|d)$, the probability of topic z_w existing in tag document d. Such conditional probabilities will be further used to predict the inter chunklet variation and intra chunklet variation. We discuss the generating process of the graphical model below.

Firstly, θ is the parameter for the topic distribution, which follows a Dirichlet distribution with parameter α:

$$\theta|\alpha \sim Dir(\alpha) \tag{4}$$

Further, given θ, topic z is drawn from a multinomial distribution, and Φ_a and Φ_w follow some Dirichlet distributions:

$$z|\theta \sim Multi(\theta), \quad \Phi_a|\beta_a \sim Dir(\beta_a), \quad \Phi_w|\beta_w \sim Dir(\beta_w) \tag{5}$$

Here we denote $\beta = [\beta_a, \beta_w]$. Finally, given topic z, both tags and visual words follow multinomial distributions:

$$w|z_w, \Phi_w \sim Multi(\phi_w^z), \quad a|z_a, \Phi_a \sim Multi(\phi_a^z) \tag{6}$$

3.2.2 *Inferences.* The main idea of the graphical model is to capture the conditional joint probability of tag document d and image x. A tag document is modeled

by a bag of words $d = \{w\}$, and the image x is represented by a bag of visual words $x = \{a\}$. The joint probability $P(z, x, d|\alpha, \beta)$ can be written as:

$$P(z, x, d|\alpha, \beta) = \prod_{a,w} P(z, a, w|\alpha, \beta) = \prod_{a,w} \int_{\theta} P(z, a, w, \theta|\alpha, \beta) d\theta$$

where a represents a visual word in the social image, and w represents one of the tags with the social image. Further, according to the assumptions, the conditional joint probability of topic z, visual word a, tag w with respect to parameters α, β can be expressed as follows:

$$P(z, a, w, \theta|\alpha, \beta_a, \beta_w) \propto P(w|z_w, \Phi_w) P(a|z_a, \Phi_a) P(z|\theta) P(\Phi_a|\beta_a) P(\Phi_w|\beta_w)$$

To calculate the chain of conditional probability in the above equation, Gibbs sampling is adopted. Although variational methods can also be used, we choose the Gibbs sampling for its simplicity and applicability to our problem. Specifically, it repeatedly draws a topic z with respect to the conditional distribution. Then visual words and tags are generated with the conditional probability given the topic z.

The objective of inference in the Gibbs sampling is to obtain the conditional distribution of hidden topic given the observed data. The Bayesian estimation of conditional distributions of tag, visual words, and topics are calculated as:

$$\begin{aligned} P(z_{w,i} = j|w) &= \frac{n^w_{-i,j} + \beta_w}{n^{\cdot}_{-i,j} + W\beta_w}, \quad P(z_{a,i} = j|a) = \frac{n^a_{-i,j} + \beta_a}{n^{\cdot}_{-i,j} + A\beta_a} \\ P(x|z_{a,i} = j) &= \frac{n^x_{-i,j} + \alpha}{n^x_{-i,\cdot} + m\alpha}, \qquad P(d|z_{w,i} = j) = \frac{n^d_{-i,j} + \alpha}{n^d_{-i,\cdot} + m\alpha} \end{aligned}$$

In the above, $z_{w,i}$ represents topic z for tag w in the i^{th} sampling; $z_{a,i}$ denotes topic z for visual word a in the i^{th} sampling; $n^w_{-i,j}$ is the frequency of tag w assigned to the j^{th} topic before the i^{th} sampling; $n^{\cdot}_{-i,j}$ is the number of all tags/visual words assigned to the j^{th} topic before the i^{th} sampling; $n^a_{-i,j}$ is the frequency of visual word a assigned to the j^{th} topic before the i^{th} sampling; $n^x_{-i,j}$ is the frequency of the j^{th} topic that appears in image x before the i^{th} sampling; $n^d_{-i,j}$ is the frequency of the j^{th} topic that appears in tag document d before the i^{th} sampling. Besides, W is the size of the tag dictionary, A is the size of the visual word dictionary, and m is the number of topics.

With the above estimations, we can calculate the marginal by integrating out the parameter θ and sampling the topic with the distribution below:

$$\begin{aligned} P(z_{w,i} = j|z_{w,-i}, w) &\propto \frac{n^w_{-i,j} + \beta_w}{n^{\cdot}_{-i,j} + W\beta_w} \times \frac{n^d_{-i,j} + \alpha}{n^d_{-i,\cdot} + m\alpha} \\ P(z_{a,i} = j|z_{a,-i}, a) &\propto \frac{n^a_{-i,j} + \beta_a}{n^{\cdot}_{-i,j} + A\beta_a} \times \frac{n^x_{-i,j} + \alpha}{n^x_{-i,\cdot} + m\alpha} \end{aligned}$$

Finally, we can calculate the topic relationship given parameter α and β below:

$$P(z_i, z_j|\alpha, \beta) \propto \frac{1}{N^2} \sum_{k=1}^{N} P(z_i, x_k, d_k|\alpha, \beta) P(z_j, x_k, d_k|\alpha, \beta)$$

where z_i and z_j are any two topics from the set Z.

As a summary, each topic z_i represents a chunklet. we can compute the conditional probability $P(z_i|x,d)$ that represents the relationship between the example and the chunklet, and the joint probability $P(z_i,z_j|\alpha,\beta)$ that represents the relationship between the two chunklets. These probabilities can be adopted and explored for DML.

3.3 Generating Chunklets by Clustering

Besides the complex topic model approach, it is also possible to study other methods to generate the probabilistic chunklets as long as the technique is able to find out the probability relationship between the examples. Below we discuss another approach, the fuzzy k-means (FKM) clustering method [Bezdek 1981], for generating the latent chunklets.

The fuzzy k-means clustering algorithm [Bezdek 1981] partitions a set of n data samples $x_1, x_2, \cdots, x_n$ into k clusters such that the overall distances of examples within the same clusters are minimized. Specifically, the optimization task of FKM can be formulated as follows:

$$\min_{P,C} J(P,C) = \sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij}^{\phi} d_{ij}^2 \tag{7}$$

$$s.t. \quad \sum_{j=1}^{k} p_{ij} = 1, i = 1,\ldots,n;$$

$$0 \le p_{ij} \le 1, i = 1,\ldots,n,\ j = 1,\ldots,k. \tag{8}$$

where $P \in \mathbb{R}^{n\times k}$ is the membership matrix, whose element $p_{ij} \in [0,1]$ indicates the probability of each data point belonging to each of the clusters (chunklets). $C = [c_1, \cdots, c_m]$ denote the centroids of the clusters (chunklets). The exponent ϕ is the fuzzy exponent which determines the degree of fuzziness, and d_{ij} is the distance between the i^{th} example and the j^{th} cluster/chunklet:

$$d_{ij}^2 = (x_i - c_j)^\top \mathbf{M}(x_i - c_j) \tag{9}$$

where x_i denotes the features of the i^{th} image, and $\mathbf{M}$ is the distance metric. If $\mathbf{M}$ is equal to an identity matrix, the distance measure reduces to Euclidian space. Here we use the tag vector to represent each image. Each image is represented as a K-dimensional vector, and the k^{th} dimension of x_{ik} indicates whether the image contains the k^{th} tag, i.e., if the k^{th} tag appears in the i^{th} image, $x_{ik} = 1$; otherwise $x_{ik} = 0$.

By clustering the social images based on the tag vectors using the FKM algorithm, we can achieve the clustering results, which include both the set of clusters/chunklets and the membership matrix P that describes the assignment probability of each example to the chunklets. Such output membership matrix P will then be used as the probabilistic chunklets in the subsequent PDML task.

4. PROBABILISTIC DISTANCE METRIC LEARNING

4.1 Problem Definition

In this section, we present a probabilistic DML (PDML) method for learning metrics from probabilistic side information. Unlike regular RCA learning, the latent chunklets are represented by some probabilistic distributions rather than "strictly-hard" pairwise constraints. Therefore, the challenge of PDML is how to exploit the uncertain side information for optimizing the metric in the most effective way. Below we present a probabilistic RCA technique, which extends the regular RCA in a probabilistic metric learning approach. We first introduce some definitions and notations below.

Let us denote by x_i a d-dimensional visual feature vector of an image, and z_k one of m latent chunklets. Further, we denote by μ_k a center (mean) for a latent chunklet z_k, and $\mu = (\mu_1, \ldots, \mu_m)$ a matrix of all centers. Moreover, we denote by matrix $P = (p_1, \ldots, p_n)$ the membership probabilities of associating examples with chunklets, where $p_i = (p_i^{(1)}, \ldots, p_i^{(m)})$ is the probability distribution for the i^{th} example and $p_i^{(k)}$ represents the probability of observing example x_i given chunklet z_k, i.e., $p_i^{(k)} = p(x_i|z_k)$.

In our approach, we initialize P by a prior probability matrix $P_0 = [p(x_i|z_k)]_{n\times m}$, which were obtained from the Latent Chunklet Estimation or the clustering process.

4.2 Probabilistic Relevant Component Analysis

The objective of our DML task is to learn an optimal metric M in a d-dimensional feature vector space, i.e., $M \in \mathbb{R}^{d\times d}$. To exploit latent chunklets in DML, we formulate a probabilistic extension of RCA, termed as "Probabilistic Relevance Component Analysis" (pRCA), as follows:

$$\min_{M\succeq 0,\mu,P} \quad \sum_{i=1}^{n}\sum_{k=1}^{m} p_i^{(k)} \|x_i - \mu_k\|_M^2 - \lambda \log|M| \tag{10}$$

$$s.t. \quad \|P - P_0\|_F^2 \leq \gamma, \tag{11}$$

$$\sum_k p_i^{(k)} = 1, p_i^{(k)} \geq 0, i = 1, \ldots, n \tag{12}$$

where parameter $\gamma \geq 0$ constraints the difference between the prior probability matrix P_0 (known from the previous side information generation stage) and the proxy probability matrix P (unknown), λ is a regularization constant, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, and $\|\cdot\|_M$ denotes the mahalanobis distance under metric M.

The above formulation can be interpreted as a robust optimization problem with bounded uncertainty on the probability matrix P. In particular, for the objective function, the first term is to minimize the sum of squared distances from examples to their chunklet centers, and the second term is to prevent the solution M from being obtained by shrinking the entire solution space. For the constraints, the one in (11) is to restrict the matrix of desired probability assignments P to be close to the prior matrix P_0, and the remaining set of constraints in (12) are used to enforce the probability requirements. The following corollary shows that RCA can be viewed as a special case of pRCA.

COROLLARY 1. *For the optimization in (10), when fixing the means of chunklets μ and the matrix of probability assignments P (assuming with hard assignments of 0 and 1), the pRCA formulation reduces to regular RCA learning.*

The proof of Corollary 1 can be found in Appendix A.

We now discuss techniques to solve the optimization of pRCA. Generally, the problem in (10) is a nonlinear optimization task containing three sets of variables M, P, and μ, where μ can be easily computed once P is found. It is often hard to solve the problem with global optima directly. To address this challenge, we present an iterative optimization algorithm by applying alternating optimization techniques [Bezdek and Hathaway 2003], which is widely used to solve multi-variable nonlinear optimization tasks.

Our iterative optimization algorithm consists of three steps: (1) fixing P and μ to optimize M; (2) fixing M and μ to optimize P; and (3) fixing P and M to find μ. According to Corollary 1, the first step is equivalent to solving regular RCA, i.e., $M = \frac{1}{\lambda}\tilde{C}^{-1}$, where $\tilde{C}$ is the average chunklet covariance matrix with the given P. The last step is straightforward, i.e., $\mu = P^\top X$, where X is a matrix of all training data.

We now focus on the second step. In particular, by fixing M and μ, the optimization can be rewritten as follows:

$$\begin{aligned} \min_{P} \quad & \sum_{i=1}^{n}\sum_{k=1}^{m} p_i^{(k)} \|x_i - \mu_k\|_M^2 + \frac{\gamma}{2}\|P - P_0\|_F^2 \\ s.t. \quad & \sum_{k} p_i^{(k)} = 1, p_i^{(k)} \geq 0, i = 1, \ldots, n \end{aligned} \tag{13}$$

where the constraint in (11) was moved to the objective. The above problem is a quadratic program (QP), which can be solved by some existing convex optimization software. However, for a real web application, the training data size can be very large, this poses a challenge of huge computation when solving a large-scale QP problem by a standard QP solver. To this end, we develop a fast algorithm, which is able to solve the above optimization very efficiently.

To ease our discussion, we notice that $p_i, i = 1, \ldots, n$ are completely decoupled in (13) given μ_k. Thus, we can rewrite (13) into a set of n independent optimization tasks, one for each p_i, i.e.,

$$\begin{aligned} \min_{p \in \mathbb{R}^m} \quad & \sum_{k=1}^{m} p_k \|x_i - \mu_k\|_M^2 + \frac{\gamma}{2}\|p - p_0\|_2^2 \\ s.t. \quad & \sum_{k=1}^{m} p_k = 1, p_k \geq 0, k = 1, \ldots, m \end{aligned} \tag{14}$$

It can be easily shown that solving the above problem is equivalent to solving the problem in (13). We now discuss a fast algorithm to solve this problem. We first introduce the Lagrangian of the optimization as follows:

$$\mathcal{L} = f^\top p + \frac{\gamma}{2}\|p - p_0\|_2^2 + \rho\Big(\sum_k p_k - 1\Big) - \eta \cdot p \tag{15}$$

where $f^\top = (\|x_i - \mu_1\|_M^2, \ldots, \|x_i - \mu_m\|_M^2)$, ρ is a Lagrange multiplier and η is a

Algorithm 1 Probabilistic RCA Algorithm (pRCA)

INPUT:
  training data matrix: $X \in \mathbb{R}^{n\times d}$
  chunklet assignment probabilities: $P_0 \in \mathbb{R}^{n\times m}$
  penalty parameter: $\gamma \geq 0$
OUTPUT:
  optimized distance metric: M^*
initialize $P = P_0$, and $\mu = P^\top X$
repeat
  (1) compute M by the following formula:
  $M = \left(\sum_{k=1}^{m}\sum_{i=1}^{n} p_i^k (x_i - \mu_k)(x_i - \mu_k)^\top\right)^{-1}$
  (2) find P by solving QP problem in (13) as follows:
  for $i = 1$ *to* n **do**
    $\mathbf{f}^\top = (\|x_i - \mu_1\|_M^2, \ldots, \|x_i - \mu_m\|_M^2)$
    $\mathbf{f} = \mathbf{sort}(\mathbf{f}, \text{'}descending\text{'})$
    find ρ by Proposition 1
    for $k = 1$ *to* m **do**
      $p_i^{(k)} = \max\left(0, p_{0k} - \frac{1}{\gamma}(\rho + f_k)\right)$
    end for
  end for
  (3) update the chunklet means: $\mu = P^\top X$
until convergence

vector of non-negative Lagrange multipliers. By differentiating it with respect to p_k, we can get the following optimality condition:

$$\frac{\partial \mathcal{L}}{\partial p_k} = f_k + \gamma(p_k - p_{0k}) + \rho - \eta_k = 0$$

By applying the KKT condition, whenever $p_k > 0$, η_k should be zero. Therefore, if $p_k > 0$, we have the following result:

$$p_k = p_{0k} - \frac{1}{\gamma}(\rho + f_k)$$

Combining the fact that $p_k \geq 0$, we have the following:

$$p_k = \max\left(0, p_{0k} - \frac{1}{\gamma}(\rho + f_k)\right) \tag{16}$$

The next issue is to find the optimal ρ. The following proposition provides a solution to find the optimal value of ρ by a simple sorting approach[Wu et al. 2009].

We can solve the QP problem (14) in $\mathcal{O}(n\log(n))$, which is significantly faster than standard QP solvers with interior point methods that usually require $\mathcal{O}(n^3)$ complexity. Finally, we summarize the pseudo-code of the pRCA algorithm in Algorithm 1. The following corollary guarantees the convergence of the proposed algorithm.

COROLLARY 2. *Algorithm 1 converges to the local optimum for the optimization problem of probabilistic relevance component analysis in (10).*

Algorithm 2 Linear projection method for searching P.

INPUT:
  a matrix p_{0k} and the scaler distance $f_k > 0$
INITIALIZE:
  $U = [n], s = 0, \rho = 0$
OUTPUT:
  $p_i^{(k)} = \max\left(0, p_{0k} - \frac{1}{\gamma}(\rho + f_k)\right)$
repeat
  Pick $k \in U$ at random;
  Partition U:
    $G = \{j \in U | p_j > p_k\}$
    $L = \{j \in U | p_j < p_k\}$
  Calculate $\triangle\ \gamma = |G|$, $\triangle\ \rho = \sum_{j \in G} p_j$
  IF $(\rho + \triangle\ \rho) - (\gamma + \triangle\ \gamma) < f_k$
  $\rho \leftarrow \rho + \triangle\ \rho; \gamma \leftarrow \gamma + \triangle\ \gamma; U \leftarrow L$
  ELSE
  $U \leftarrow G \backslash \{k\}$
  ENDIF
until $U = \varnothing$

4.3 Probabilistic Discriminative Component Analysis

Similarly, we can also generalize the DCA technique [Hoi et al. 2006] by applying the proposed probabilistic distance metric learning framework in order to incorporate both positive and negative pairwise constraints. Specifically, we formulate the probabilistic Discriminative Component Analysis method (pDCA) as follows:

$$\min_{M \succeq 0, \mu, P} \quad \sum_{i=1}^{n} \sum_{k=1}^{m} p_i^{(k)} \|x_i - \mu_k\|_M^2 + \gamma \|P - P_0\|_F^2 \tag{17}$$

$$s.t. \quad \sum_{i \neq j} (1 - p_{ij}) \|\mu_i - \mu_j\|_M^2 \geq 1 \tag{18}$$

$$\mu_k = \sum_{i=1}^{n} p_i^{(k)} \mathbf{x}_i, k = 1, \ldots, m, \quad \sum_k p_i^{(k)} = 1, p_i^{(k)} \geq 0, i = 1, \ldots, n$$

where p_{ij} denotes the joint probability of two web images, which is estimated by the latent chunklet estimation process, e.g., $p_{ij} = P(z_i, z_j | \alpha, \beta)$ in the graph model approach. As a result, $(1 - p_{ij})$ measures the dissimilarity between any two chunklets, which implicitly represents the probability of negative constraint. Therefore, in the above formulation, the first constraint is introduced to avoid two dissimilar chunklets from being too close by exploring negative constraints.

Similar to the approach in solving pRCA, we can also solve the pDCA problem by an iterative algorithm of three steps. The first step is to fix P and μ and then

optimize M, for which the optimization can be reduced as follows:

$$\min_{M \succeq 0} \quad \sum_{i=1}^{n} \sum_{k=1}^{m} p_i^{(k)} \|x_i - \mu_k\|_M^2 \tag{19}$$

$$s.t. \quad \sum_{i \neq j} (1 - p_{ij}) \|\mu_i - \mu_j\|_M^2 \geq 1 \tag{20}$$

It can be shown that the above optimization is almost equivalent to the regular DCA. The second step of the iterative algorithm is to fix M and μ, and then optimize P. For this step, it is clear to see that the reduced optimization of the pDCA formulation in (17) becomes the same QP problem as shown in (13). The last step is to update the chunklet means μ based on the optimized P. Finally, Algorithm 3 summarizes the iterative algorithm of probabilistic DCA.

Algorithm 3 Probabilistic DCA Algorithm (pDCA)

1: INPUT:
   training data matrix: $X \in \mathbb{R}^{n \times d}$
   chunklet assignment probabilities: $P_0 \in \mathbb{R}^{n \times m}$
   chunklet joint probabilities: $\hat{P} \in \mathbb{R}^{m \times m}$
   penalty parameter: $\gamma \geq 0$
2: OUTPUT:
   optimized distance metric: M^*
   proxy probabilities of chunklet assignments: P^*
3: initialize $P = P_0$
4: initialize $\mu = P^\top X$
5: **repeat**
6: (1) compute M by solving DCA optimization in (19)
7: (2) find P by solving the QP optimization in (13)
8: (3) update the chunklet means: $\mu = P^\top X$
9: **until** convergence

5. APPLICATION TO AUTOMATED PHOTO TAGGING

In this section, we discuss the application of pRCA to the exploitation of social photo repositories for automated photo tagging tasks. Given a novel photo, the automated tagging task is to annotate the photo labels or tags, which often reflect certain semantic concepts/objects. To overcome the limitation of conventional approaches, we investigate a retrieval based approach to automated photo tagging tasks by exploring a huge number of social photos freely available on the web. We formally formulate our approach as follows.

Let $I_q = \{x_q, \mathcal{T}_q\}$ denote a query image for tagging, where x_q represents the visual contents of the image, and $\mathcal{T}_q$ denotes a set of unknown tags to be found in the tagging task. In general, a retrieval based tagging approach consists of two steps: (1) retrieving a set of visually similar social photos, which are closest to the query photo; and (2) annotating the query photo by a set of most relevant tags that are associated with the retrieved similar photos.

For the first step, there are two typical approaches to find a set of nearest neighbors with respect to a query image. One is to retrieve the k-nearest neighbors of the query image, i.e.,

$$\mathcal{N}_k(x_q) = \{i \in [1, \ldots, n] | x_i \in \text{kNN list}(x_q)\} \,, \tag{21}$$

where n is the total number of photos in the social photo repository. The other way is to retrieve a set of nearest photos within certain distance range, i.e.,

$$\mathcal{N}_\epsilon(x_q) = \{i \in [1, \ldots, n] | \ \|x_i - x_q\|_M \leq \epsilon\} \,, \tag{22}$$

where ϵ is a predefined distance threshold. For both approaches, it is clear that an effective distance metric M is essential to retrieve the set of nearest neighbors. In this paper, we adopt the first approach and employ the metric learned by pRCA to compute the k-NN list.

For the second step, we suggest an information theory based tag ranking scheme by adopting the voting by maximum likelihood scheme. Specifically, we define a set of candidate tags $\mathcal{T}_w$ as:

$$\mathcal{T}_w = \bigcup_{i \in \mathcal{N}_k} \mathcal{T}_i \tag{23}$$

where $\mathcal{T}_i$ represents the set of tags associated with image I_i. For each candidate tag $w_j \in \mathcal{T}_{w_j}$, we compute its frequency appearing in the k^{th} nearest web photos I_k, denoted by $f(w_j|I_k)$. The conditional probability of each tag given the k^{th} similar photo I_k is calculated as follows.

$$p(w_j|I_k) = \frac{f(w_j|I_k) + 1}{\sum_l f(w_l|I_k) + \kappa}$$

where κ is a smoothing parameter which is simply fixed to the vocabulary size in our experiments. The likelihood of assigning the tag w_j to the test image I_i is

$$p(w_j|I_i) = \sum_k p(w_j|I_k) p(I_k|I_i)$$

where $p(I_k|I_i)$ is estimated by the visual similarity between two images, which is calculated by

$$p(I_k|I_i) = exp(-\gamma \|I_k - I_i\|_M)$$

where we use a Gaussian kernel to model the visual conditional probability and γ is a kernel parameter that is empirically determined by a validation set.

We then incrementally add the best tag w^* into the tag set for the query image $\mathcal{T}_q = \mathcal{T}_q \cup \{w^*\}$, which is chosen according to their likelihood scores, i.e.,

$$w^* = \underset{w \in \mathcal{T}_w \wedge w \notin \mathcal{T}_q}{\arg\max} \ p(w|I_i) \tag{24}$$

where $p(w|I_i)$ represents the probability the candidate photo I_i that contains tag w. The above formula indicates that we prefer to assign the query image with a tag according to both tag frequency and image visual similarity.

6. EXPERIMENTS

The goal of our experiment is to examine if the proposed distance metric learning method is more effective than conventional methods for automated photo tagging tasks. To this purpose, we first conduct a numerical evaluation by comparing the proposed algorithms with a number of state-of-the-art distance metric learning algorithms, and further examine the influence of varied parameters and settings that could affect the performance of the proposed automated photo tagging scheme. Finally, we note that all experiments were run in the same environment with a typical PC of 2.8GHz CPU with Matlab.

6.1 Experimental Testbed

We collected a large social photo testbed with over 1,000,000 photos crawled from www.Flickr.com, in which most photos contain user-tags and other metadata. There are around 200,000 tags in the dataset. The average occurrence of each tag is around 11. We split the whole dataset into three disjoint partitions: a *training* set, a *test* set, and a *knowledge database* set. Since both the images for metric learning and for knowledge databased are crawled from Flickr, the tag property and distribution of the two sets are similar. Below we describe the details of the three partitions.

The training set is used for learning distance metrics. In particular, we randomly sampled 16,588 photos with tags from the whole social photo testbed. We did not make any refinements on the associated tags. To provide visual words for training the models, we construct the bag-of-visual-words representation by extracting local features from the training photos using the SIFT descriptor [Lowe 2004].

The test set is used for evaluating the photo tagging performance. In particular, we randomly picked 2,000 photos from the whole photo testbed as the query images to test the photo tagging performance. To improve the quality of test data, we created the annotation ground truth by manually removing some clear noises to refine the original tags.

The rest social photos are engaged as the knowledge database set, which serves the base of social photo repository for tagging. We also randomly selected 200,000 photos from the knowledge database. We perform directly similarity search on this small knowledge database, for the comparison with the search results in the whole knowledge database, in which LSH indexing [Andoni and Indyk 2008] is adopted to improve the search efficiency. We try to see whether the scale of the knowledge database will help improve the performance.

Finally, for the photos in both test set and the knowledge database, we extract a set of effective and compact visual features [Hoi et al. 2006; Hoi et al. 2009], including: (1) grid color moments, (2) edge direction histogram, (3) Gabor textual features, and (4) Local binary pattern histograms. In total, a 297-dimensional feature vector is used to represent each photo. The reason that here we do not adopt local features, such as SIFT, is primarily due to the efficiency consideration.

6.2 Compared Schemes

To examine the effectiveness of our technique, we compare the proposed pRCA and pDCA algorithm with some baseline and a number of state-of-the-art DML meth-

ods, including (1) a **baseline** that simply adopts Euclidean distance, (3) regular RCA [Bar-Hillel et al. 2005], (3) Discriminative Component Analysis (DCA) [Hoi et al. 2006], (4) Information-Theoretic Metric Learning (ITML) [Davis et al. 2007], (5) Large Margin Nearest Neighbor (LMNN) [Weinberger et al. 2006], (6) Neighbourhood Components Analysis (NCA) [Goldberger et al. 2005], and (7) Regularized Distance Metric Learning (RDML) [Si et al. 2006]. Note that we excluded other DML methods in our comparison mainly due to their computational infeasibility for such large-scale applications. For example, the well-known DML method in [Xing et al. 2002] is only applicable to a very small dataset.

Regarding the two proposed algorithms, pRCA and pDCA, there are some common property, i.e., both of them adopt the probabilistic constraints, which is also the key advantage over traditional RCA and DCA methods. In general, pDCA can be viewed as an extension of pRCA. The difference is that pRCA only minimizes the distance between the relevant samples, while pDCA both minimize the distance between high relevant samples and maximize the distance between low relevant samples.

Since no explicit side information is available for traditional DML, in training stage, we performed clustering on training photos using both visual features and tag co-occurrence information. Photos that have similar visual contents and share common tags will be grouped together. Finally, we generate side information from the resulting clusters (after removing trivial clusters) as the inputs for DML.

We sample the same subset of image pairs for both deterministic metric learning and probabilistic metric learning. For the probabilistic metric learning, we estimate the probabilistic chunklets by the sample image content and their tags. For the deterministic metric learning, if the sampled pair of images share any tag, they are in same chunk; otherwise, in different chunks

6.3 Experimental Setup and Protocols

Regarding parameter settings, for the pRCA learning, we assume there are m ($m = 500$) latent chunklets for the N ($N = 16,588$) training examples, and generate an $m \times N$ matrix of probabilistic latent chunklets distribution by the graphical model as the probabilistic side information, which is used as the prior probability matrix P_0 for metric learning. For the extraction of visual words in LCE, we set the number of visual words $A = 1,000$, and the number of tags $W = 2,000$. The parameter γ of pRCA was simply fixed to 0.5 for all experiments.

For other DML methods, we adopt the same settings, i.e., 500 chunklets for producing the side information. For their parameters, we chosen them according to the suggestions/empirical results in the original work.

To evaluate the automated photo tagging performance by different methods, we employ the proposed retrieval-based annotation solution presented in Section 5. Firstly, for each query photo in the test set, top k nearest photos from the database are first retrieved as the set of candidate images. Then, we annotate the query photo by assigning top t tags ranked by the function in (24). Finally, we adopt standard average precision and average recall at top t tags as performance metrics to evaluate the automated photo tagging performance.

6.4 Experiment I: Numerical Evaluation

Figure 3 and Figure 4 show average precision and average recall at top t annotated tags, respectively. For these results, we fixed the number of nearest neighbors k to 30 for all compared methods. In both figures, the horizontal axis denotes the number of top tags t that ranges from 1 to 10.

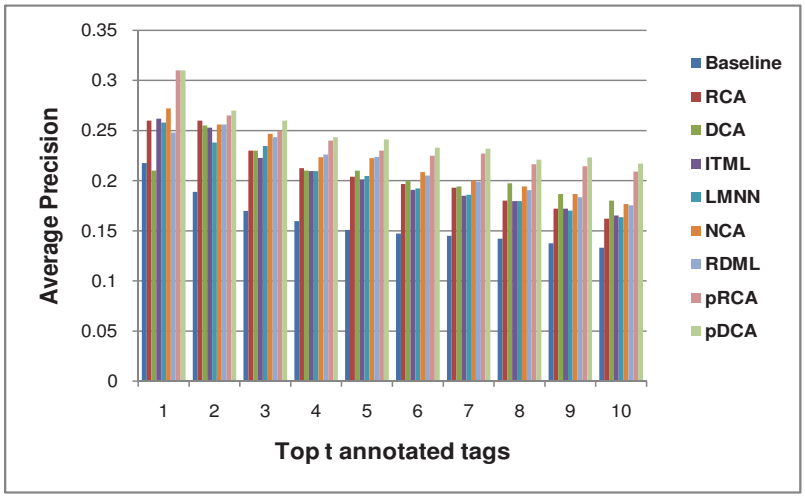


Fig. 3. Average precision at top t annotated tags

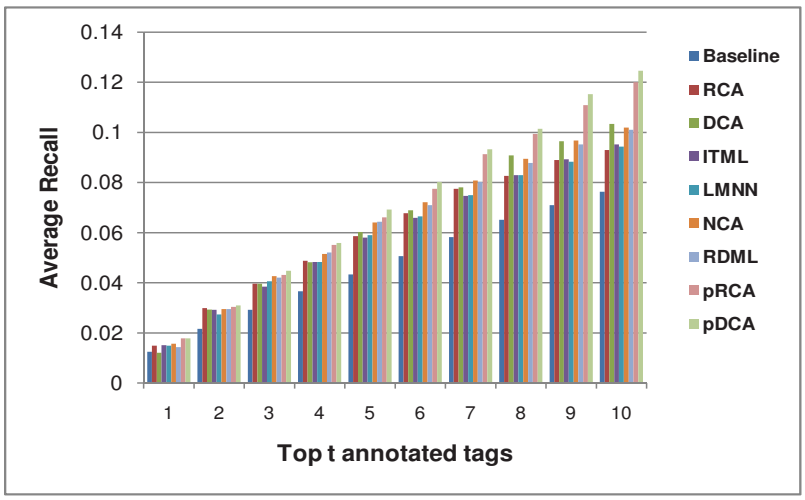


Fig. 4. Average recall at top t annotated tags

From the figures, we can draw several observations. First of all, we found that most DML techniques outperformed the baseline by simple Euclidean distance. This shows that DML techniques are beneficial and critical to the retrieval-based photo tagging tasks. Second, we found that for some cases, some DML methods

did not perform well, and sometimes performed even worse than the Euclidean method. For example, for the case of top-1 annotated tag, we found that DCA performed slightly worse than Euclidean. We believe this is mainly due to the noisy side information issue. This again shows that it is important to develop some effective and robust method in our problem. Further, we observe that the proposed pRCA algorithm considerably outperformed other approaches in most cases. For instance, for the case of top-1 tag, pRCA achieved average precision of about 31%, which improves the baseline approach over 40% and over RCA about 20%. Finally, comparing the two proposed methods, pRCA and pDCA, we found they are quite comparable, in which pDCA tends to be slightly more effective than pRCA.

Figure 5 further shows the precision-recall curves. Similar observations were found. The proposed algorithms, pRCA and pDCA, considerably outperform the others. This is because our methods use the probabilistic constraints rather than the traditional hard constraints. The probabilistic constraints can better reflect the relationship between the examples and thus achieve more accurate results. These results again validate the efficacy and significance of our technique.

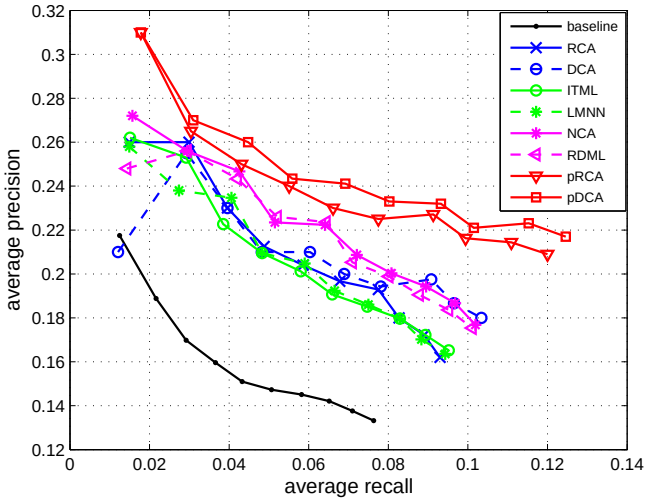


Fig. 5. Comparison of the precision-recall curves

6.5 Experiment II: Evaluation of Varied k Values

We also notice that an important parameter, i.e., the number of nearest neighbors k, could affect the annotation performance considerably. To examine how is its impact, we evaluate the annotation performance of the proposed annotation method by varying the value of parameter k. Figure 6 and 7 show the average precision results of the proposed pRCA and pDCA annotation approaches by varying the value of k from 10 to 50.

From the experimental results, we found that when k equals to 30, the resulting annotation performance is generally better than the other cases. We suspect the main reason is that if we set k too large, e.g. 50, many noisy tags may be included; as a result, there may not exist so many relevant images in the database, which

thus could harm the performance. However, if we set k too small, some relevant tags may not appear, which again would degrade the annotation performance.

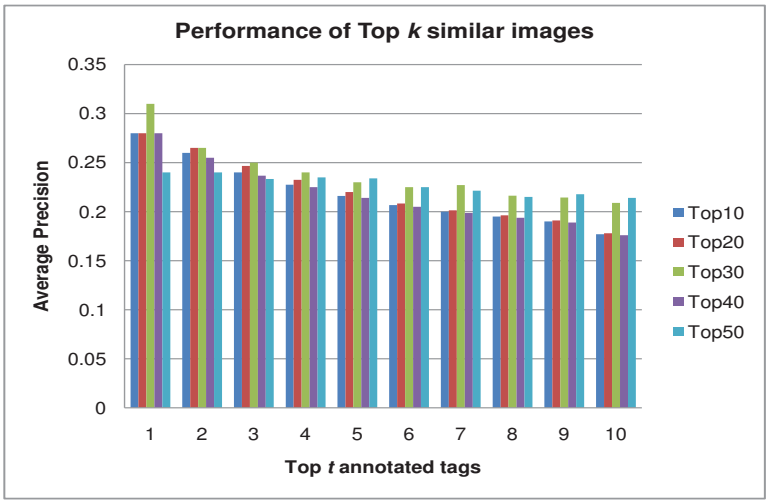


Fig. 6. Average precision at top t tags using top k retrieved images by pRCA for annotation.

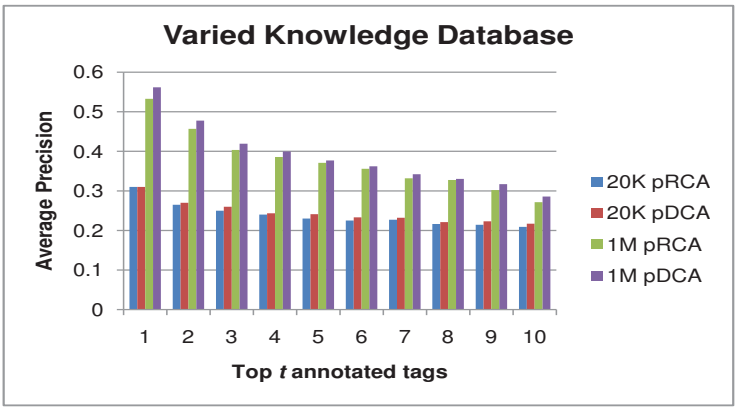
Fig. 7. Average precision at top t tags using top k retrieved images by pDCA for annotation.

6.6 Experiment III: Influence of the Knowledge Database Sizes

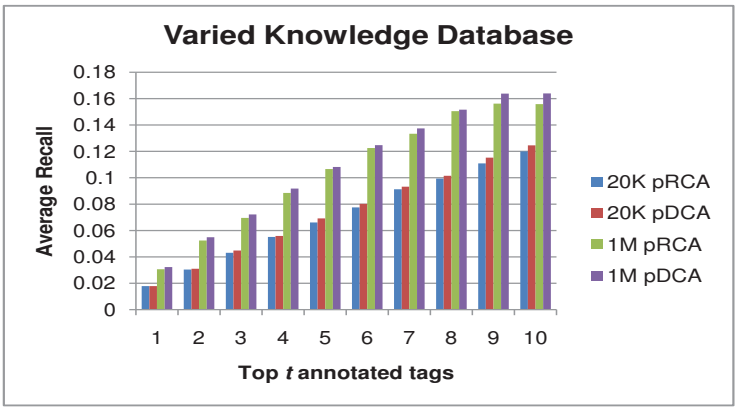
In our annotation framework, the size of the knowledge database plays a critical role in affecting the annotation quality. In this experiment, we aim to evaluate how the size of knowledge database affects the image annotation performance. In particular, we vary the size of the knowledge database from 20,000 to 1,000,000 and

evaluate the average precision/recall of image annotation based on each database. Fig. 8 summarizes the comparison of the annotation performance with respect to two knowledge databases of different sizes.

As we can see from the figure, when the size of the knowledge database increases, the performance of the retrieved based photo tagging solution is improved considerably. The main reason is because once we have a larger database, the chance of finding the similar/relevant images can be potentially increased, which thus leads to the improvement of the annotation quality as the performance of retrieval-based tagging method highly depends on the relevance of the retrieved similar images.



(a) Average Precision.



(b) Average Recall.

Fig. 8. Performance of image annotation with different knowledge databases.

6.7 Experiment III: Time Cost Evaluation

The third experiment is to evaluate the time efficiency of the proposed DML algorithm. To this purpose, we compare time performance of our algorithm with other

DML algorithms. Table I summarizes the evaluation of average time cost results that were obtained by running the compared algorithms in our DML tasks.

Table I. Time cost comparison of different DML methods (seconds).

(s)	baseline	RCA	DCA	ITML	LMNN	NCA	RDML	pRCA	pDCA
Time	N/A	731.6	865.6	1185.3	1673.2	28989.8	824.8	891.2	936.5

From the results, we can see that the most efficient method is the regular RCA approach, and the least efficient one is NCA that was significantly slower than the others. Finally, by comparing our algorithms with the other competing algorithms, we found that both pRCA and pDCA are quite competitive, which though are slightly worse than RCA, DCA, and RDML, are considerably more efficient than ITML, LMNN, and NCA.This is due to the efficient sorting algorithm. Since we use a sorting algorithm instead of to solving the QP problem directly, our methods can be much faster than its counterparts.

6.8 Experiment IV: Generating Latent Chunklets: Sampling vs. Clustering

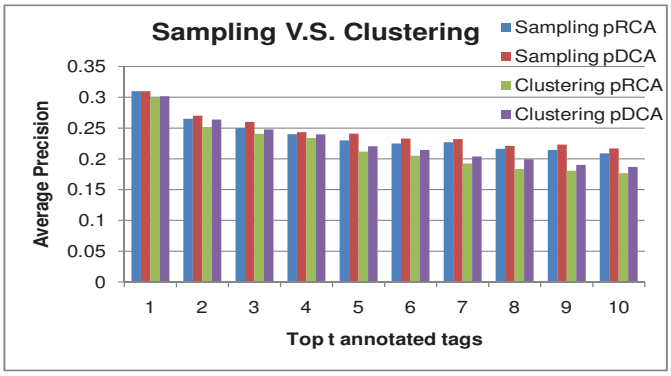
As discussed previously, we suggest two kinds of approaches to generating the latent chunklets (i.e., side information). One is the sampling method using the graphical model, and the other is the clustering approach using the fuzzy k-means. In this section, we aim to compare the sampling method with the clustering method to examine their influence on the final image annotation task.

We evaluate the performance of both methods by computing their average precision and average recall scores. For the clustering based approach, we adopt the fuzzy k-means algorithm [Bezdek 1981], which also generates a soft probabilistic relationship between samples and the clusters. For fair comparison, we generate the same numbers of chucklets/clusters using the same settings for both compared methods. Fig. 9 shows the results of average precision and average recall of the image annotation task.

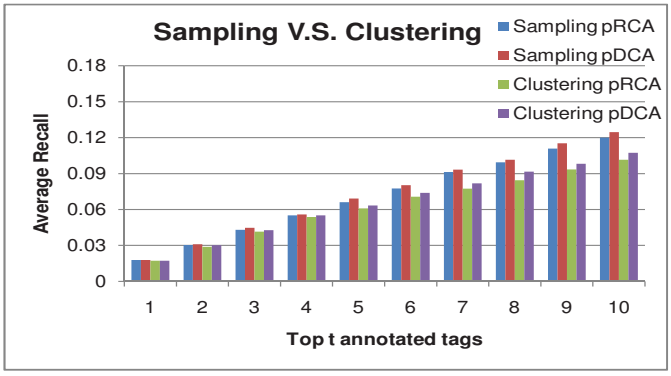
From the experimental results, we found that both methods perform quite comparably for the automated image tagging task. Empirically, the graphical model based approach is slightly better than the clustering based approach. This is reasonable as the graphical model may generate more natural and effective initial chunklets compared with the clustering based approach. Since the probabilistic chunklets will be automatically updated in the subsequent distance metric learning process, the initialization actually has limited influence on the final performance. This also shows that the proposed algorithm is robust to the noisy side information.

6.9 Experiment V: Qualitative Comparison

In addition to the previous quantitative evaluations, our last experiment is to examine qualitative performance of our automated photo tagging solution. We randomly picked a list of query photos from the test set and showed the qualitative retrieval and annotation results in Figure 10 and Figure 11, respectively. From these results, we can observe that our solution generally achieved better qualitative results than others. On average, our method can produce more than 5 correct annotations for each image, which is better than other methods. Also the retrieval result shows our method can produce more relevant images.



(a) Average Precision.



(b) Average Recall.

Fig. 9. Performance of image annotation with different approaches to generating side information.

7. CONCLUSIONS

This paper investigated a new problem, termed "Probabilistic Distance Metric Learning" (PDML), which aims to learn distance metrics from uncertain side information that implicitly exists in some real applications. Unlike conventional DML techniques that work with explicit side information, the PDML problem is more challenging given that the side information is not explicitly available. We proposed a novel two-stage PDML framework, which firstly discovers probabilistic side information from the data using an unsupervised learning approach, and then employs some effective probabilistic DML algorithm to find an optimal metric from the probabilistic side information. In particular, we proposed two effective PDML algorithms, i.e., probabilistic RCA and probabilistic DCA. We applied the proposed technique to automated photo tagging on a large-scale social photo testbet with over one million photos from Flickr. By comparing our technique with a number of state-of-the-art DML methods extensively, we concluded that our technique is effective and promising for solving this challenging problem. Future work will extend our framework by exploring more social information to boost automated photo tagging [Sigurbjörnsson and van Zwol 2008; Stone et al. 2008].

Acknowledgements

This paper is based on the paper "Distance Metric Learning from Uncertain Side Information with Application to Automated Photo Tagging", which appeared in proceedings of ACM International Conference on Multimedia, Beijing, P.R. China, 19-24 Oct 2009. This work was done when Lei Wu was an RA at Nanyang Technological University, Singapore. The work was supported in part by Singapore MOE Academic Tier-1 Research Grant (RG67/07), the National Science Foundation (IIS-0643494), and National Institute of Health (1R01-GM079688-01). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF and NIH.

Appendix A: Proof of Corollary 1

PROOF. By fixing μ and P, the optimization reduces to:

$$\min_{M\succeq 0} \quad \sum_{i=1}^{n}\sum_{k=1}^{m} p_i^{(k)}\|x_i-\mu_k\|_M^2 - \lambda\log|M| \tag{25}$$

By differentiating the Lagrangian with respect to M, we have the following equality:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} p_i^{(k)}(\mathbf{x}_i-\mu_k)(\mathbf{x}_i-\mu_k)^\top - \lambda M^{-1} = 0 \tag{26}$$

Hence, we have the optimal solution: $M = \frac{1}{\lambda}\hat{C}^{-1}$, where matrix $\hat{C}$ is given as:

$$\hat{C} = \sum_{i=1}^{n}\sum_{j=1}^{k} p_i^{(k)}(\mathbf{x}_i-\mu_k)(\mathbf{x}_i-\mu_k)^\top \tag{27}$$

When $p_i^{(k)}$ takes only 0 or 1, it can be seen clearly that the solution of M is almost identical to the solution learned by RCA (up to a global scale factor). Hence, pRCA reduces to regular RCA learning in this special case. □

Appendix B: Proofs of pRCA Solution

Here we discuss the details of our techniques in solving the QP problem in (18) and also give some formal proofs of our approach. We consider the optimization problem as follows:

$$\begin{aligned} \min_{\mathbf{p}\in\mathbb{R}^m} \quad & \sum_{k=1}^{m} p_k\|x_i-\mu_k\|_M^2 + \frac{\gamma}{2}\|p-p_0\|_2^2 \\ s.t. \quad & \sum_{k=1}^{m} p_k = 1, p_k \geq 0, k=1,\ldots,m \end{aligned} \tag{28}$$

We have the Lagrangian of the optimization as follows:

$$\mathcal{L} = f^\top p + \frac{\gamma}{2}\|p-p_0\|_2^2 + \rho\Big(\sum_k p_k - 1\Big) - \eta\cdot p \tag{29}$$

where $f^\top = (\|x_i-\mu_1\|_M^2,\ldots,\|x_i-\mu_m\|_M^2)$, ρ is a Lagrange multiplier and η is a vector of non-negative Lagrange multipliers. By differentiating it with respect to

Fig. 10. Examples showing top images retrieved by different methods. For each row, the first image is a test image for tagging and each following block shows top 4 images retrieved by one of the compared methods.

p_k, we have:

$$\frac{\partial \mathcal{L}}{\partial p_k} = f_k + \gamma(p_k - p_{0k}) + \rho - \eta_k = 0$$

By applying the KKT condition, whenever $p_k > 0$, η_k should be zero. Therefore, if $p_k > 0$, we have the following result:

$$p_k = p_{0k} - \frac{1}{\gamma}(\rho + f_k)$$

Fig. 11. Examples showing the tagging results by different methods. For each row, the first image is a test image and each following block shows top 10 tags annotated by one method. The correct tags are highlighted.

Combining the fact that $p_k \geq 0$, we have the following:

$$p_k = \max\left(0, p_{0k} - \frac{1}{\gamma}(\rho + f_k)\right) \tag{30}$$

To simplify the formula, we let $a_k = p_{0k} - \frac{f_k}{\gamma}$, as a result, p_k can be rewritten as:

$$p_k = \max\left(0, a_k - \frac{\rho}{\gamma}\right) \tag{31}$$

Next we show a proposition to find the optimal ρ by a simple sorting approach.

PROPOSITION 1. *Let a' denote the vector by sorting a in decreasing order, the optimal value of ρ to the solution in (31) can be computed as: $\rho = \frac{\gamma}{\tau}\Big(\sum_{k=1}^{\tau} a'_k - 1\Big)$, where τ can be found through a sorting approach, i.e.,*

$$\tau = \max_{k\in[1,n]} \Big\{k : a'_k - \frac{1}{k}\big(\sum_{j=1}^{k} a'_j - 1\big) > 0\Big\} \tag{32}$$

PROOF. In order to prove this proposition, we shall first introduce a lemma:

LEMMA 1. *Let p denote the optimal solution to the minimization problem in (3), let s and t are two indices such that $a_s > a_t$, where $a_s = p_{0s} - \frac{f_s}{\gamma}$ as defined above. If $p_s = 0$, then p_t must also be zero.*

PROOF. We can prove it by contraction, i.e., assume that $p_s = 0$ but $p_t > 0$. Let us introduce a vector p' by setting $p'_s = p_t$, $p'_t = p_s$, and $p'_k = p_k$ for $\forall k \neq s \wedge k \neq t$. It is clear the constraint $sum(p') = 1$ still holds. We now compare the two objectives:

$$obj(p) = \sum_{k=1}^{m} p_k f_k + \frac{\gamma}{2}\|p - p_0\|_2^2 \tag{33}$$

$$obj(p') = \sum_{k=1}^{m} p'_k f_k + \frac{\gamma}{2}\|p' - p_0\|_2^2 \tag{34}$$

$$obj(p) - obj(p') = \gamma p_t\Big((p_{0s} - \frac{f_s}{\gamma}) - (p_{0t} - \frac{f_t}{\gamma})\Big) = \gamma p_t(a_s - a_t) > 0 \tag{35}$$

The above result means $obj(p) > obj(p')$, which contradicts the fact that p is the optimal (minimal) solution. □

Lemma 1 implies that those non-zero solutions p_k should have the largest values of a_k. This shows we can find p_k by sorting vector a in decreasing order, denoted by a'. As a result, by combining the optimality condition, we have equation: $\sum_{k=1}^{n} p_k = \sum_{i=1}^{\tau} a'_i - \frac{\rho}{\gamma} = 1$, where τ is a constant number. Once τ is given, it is clear to have:

$$\rho = \frac{\gamma}{\tau}\Big(\sum_{k=1}^{\tau} a'_k - 1\Big) \tag{36}$$

Finally, the optimal value of τ can be found by applying the following lemma, which was proved in [Shalev-Shwartz and Singer 2006].

LEMMA 2 [SHALEV-SHWARTZ AND SINGER 2006]. *Let α be the optimal solution to the minimization problem below:*

$$\min_{\alpha} \frac{1}{2}\|\alpha - \mu\|^2, s.t. \sum_{i=1}^{n} \alpha_i = z, \alpha_i \geq 0 \tag{37}$$

and assume that μ is sorted in decreasing order. Then, the number of strictly positive elements in p is:

$$\tau = \max_{k \in [1,n]} \Big\{ k : \mu_k - \frac{1}{k}\big(\sum_{j=1}^{k} \mu_j - 1\big) > 0 \Big\} \tag{38}$$

Applying the above lemma leads to compete the proof of this proposition.

REFERENCES

Andoni, A. and Indyk, P. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM 51,* 1, 117–122.

Bar-Hillel, A., Hertz, T., Shental, N., and Weinshall, D. 2005. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research 6*, 937–965.

Bezdek, J. C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Bezdek, J. C. and Hathaway, R. J. 2003. Convergence of alternating optimization. *Neural, Parallel Sci. Comput. 11,* 4, 351–368.

Carneiro, G., Chan, A. B., Moreno, P., and Vasconcelos, N. 2006. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Tran. PAMI*, 394–410.

Carneiro, G. and Vasconcelos, N. 2005. Formulating semantic image annotation as a supervised learning problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 163–168.

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. 2007. Information-theoretic metric learning. In *International Conference on Machine Learning (ICML'07)*. 209–216.

Duygulu, P., Barnard, K., de Freitas, J., and Forsyth, D. A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*. 97–112.

Fan, J., Gao, Y., and Luo, H. 2004. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *ACM Multimedia*. 540–547.

Fukunaga, K. 1990. *Introduction to Statistical Pattern Recognition*. Elsevier.

Globerson, A. and Roweis, S. 2005. Metric learning by collapsing classes. In *NIPS'05*.

Goldberger, J., Roweis, S., Hinton, G., and Salakhutdinov, R. 2005. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*. 513–520.

He, X. and Zemel, R. S. 2008. Learning hybrid models for image annotation with partially labeled data. In *NIPS*. 625–632.

Hoi, C.-H. and Lyu, M. R. 2004. Web image learning for searching semantic concepts in image databases. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. New York, NY, USA, 406–407.

Hoi, S. C., Jin, R., Zhu, J., and Lyu, M. R. 2009. Semi-supervised svm batch mode active learning with applications to image retrieval. *ACM Transactions on Information Systems (TOIS) 27,* 3 (July), 1–29.

Hoi, S. C., Liu, W., and Chang, S.-F. 2008. Semi-supervised distance metric learning for collaborative image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*.

Hoi, S. C. H., Liu, W., Lyu, M. R., and Ma, W.-Y. 2006. Learning distance metrics with contextual constraints for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*. New York, US.

Hoi, S. C. H., Lyu, M. R., and Jin, R. 2006. A unified log-based relevance feedback scheme for image retrieval. *IEEE Transactions on Knowledge and Data Engineering (TKDE) 18,* 4, 509–204.

Jeon, J., Lavrenko, V., and Manmatha, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR'03*. Toronto, Canada, 119–126.

Jin, R., Wang, S., and Zhou, Y. 2009. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems 22*. 862–870.

Lew, M. S., Sebe, N., Djeraba, C., and Jain, R. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl. 2,* 1, 1–19.

Li, W. and Sun, M. 2006. Semi-supervised learning for image annotation based on conditional random fields. In *ACM International Conference on Image and Video Retrieval (CIVR)*. 463–472.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *IJCV 60*, 91–110.

Qi, G.-J., Hua, X.-S., and Zhang, H.-J. 2009. Learning semantic distance from community-tagged media collection. In *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*. ACM, New York, NY, USA, 243–252.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. 2008. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision 77,* 1-3, 157–173.

Shalev-Shwartz, S. and Singer, Y. 2006. Efficient learning of label ranking by soft projections onto polyhedra. *J. Mach. Learn. Res. 7*, 1567–1599.

Si, L., Jin, R., Hoi, S. C. H., and Lyu, M. R. 2006. Collaborative image retrieval via regularized metric learning. *ACM Multimedia Systems Journal 12,* 1, 34–44.

Sigurbjörnsson, B. and van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In *WWW'08: Proceeding of the 17th international conference on World Wide Web*. ACM, New York, NY, USA, 327–336.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. PAMI 22,* 12, 1349–1380.

Stone, Z., Zickler, T., and Darrell, T. 2008. Autotagging facebook: Social network context improves photo annotation. In *IEEE Workshop on Internet Vision*. IEEE.

Torralba, A., Weiss, Y., and Fergus, R. 2008. Small codes and large databases of images for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, C., Zhang, L., and Zhang, H.-J. 2008. Learning to reduce the semantic gap in web image retrieval and annotation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 355–362.

Wang, M., Zhou, X., and Chua, T.-S. 2008. Automatic image annotation via local multi-label classification. In *ACM International Conference on Image and Video Retrieval (CIVR)*. ACM, New York, NY, USA, 17–26.

Wang, X.-J., Zhang, L., Jing, F., and Ma, W.-Y. 2006. Annosearch: Image auto-annotation by search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*. 1483–1490.

Weinberger, K., Blitzer, J., and Saul, L. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems 18*. 1473–1480.

Wu, L., Hoi, S. C., Jin, R., Zhu, J., and Yu, N. 2009. Distance metric learning from uncertain side information with application to automated photo tagging. In *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*. ACM, New York, NY, USA, 135–144.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. 2002. Distance metric learning with application to clustering with side-information. In *NIPS2002*.

Yan, R., Natsev, A., and Campbell, M. 2008. A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*.

Yang, L., Jin, R., Sukthankar, R., and Liu, Y. 2006. An efficient algorithm for local distance metric learning. In *AAAI*.