Online Multi-modal Distance Metric Learning with Application to Image Retrieval

Pengcheng Wu, Steven C. H. Hoi, Peilin Zhao, Chunyan Miao, Zhi-Yong Liu

Abstract—Distance metric learning (DML) is an important technique to improve similarity search in content-based image retrieval. Despite being studied extensively, most existing DML approaches typically adopt a single-modal learning framework that learns the distance metric on either a single feature type or a combined feature space where multiple types of features are simply concatenated. Such single-modal DML methods suffer from some critical limitations: (i) some type of features may significantly dominate the others in the DML task due to diverse feature representations; and (ii) learning a distance metric on the combined high-dimensional feature space can be extremely time-consuming using the naive feature concatenation approach. To address these limitations, in this paper, we investigate a novel scheme of online multi-modal distance metric learning (OMDML), which explores a unified two-level online learning scheme: (i) it learns to optimize a distance metric on each individual feature space; and (ii) then it learns to find the optimal combination of diverse types of features. To further reduce the expensive cost of DML on high-dimensional feature space, we propose a low-rank OMDML algorithm which not only significantly reduces the computational cost but also retains highly competing or even better learning accuracy. We conduct extensive experiments to evaluate the performance of the proposed algorithms for multi-modal image retrieval, in which encouraging results validate the effectiveness of the proposed technique.

Index Terms—content-based image retrieval, multi-modal retrieval, distance metric learning, online learning

1 Introduction

One of the core research problems in multimedia retrieval is to seek an effective distance metric/function for computing similarity of two objects in content-based multimedia retrieval tasks [1], [2], [3]. Over the past decades, multimedia researchers have spent much effort in designing a variety of low-level feature representations and different distance measures [4], [5], [6]. Finding a good distance metric/function remains an open challenge for content-based multimedia retrieval tasks till now. In recent years, one promising direction to address this challenge is to explore distance metric learning (DML) [7], [8], [9] by applying machine learning techniques to optimize distance metrics from training data or side information, such as historical logs of user relevance feedback in content-based image retrieval (CBIR) systems [6], [7].

Although various DML algorithms have been proposed in literature [7], [10], [11], [12], [13], most existing DML methods in general belong to single-modal DML in that they learn a distance metric either on a single type of feature or on a combined feature space by simply concatenating multiple types of diverse features together. In a real-world application, such approaches may suffer from some practical limitations: (i) some types of features may significantly dominate the others in the DML task, weakening the ability to exploit the potential of all features; and (ii) the naive concatenation approach may result in a combined high-dimensional feature space, making the subsequent DML task computationally intensive.

To overcome the above limitations, this paper investigates a novel framework of Online Multi-modal Distance Metric Learning (OMDML), which learns distance metrics from multi-modal data or multiple types of features via an efficient and scalable online learning scheme. Unlike the above concatenation approach, the key ideas of OMDML are twofold: (i) it learns to optimize a separate distance metric for each individual modality (i.e., each type of feature space), and (ii) it learns to find an optimal combination of diverse distance metrics on multiple modalities. Moreover, OMDML takes advantages of online learning techniques for high efficiency and scalability towards large-scale learning tasks. To further reduce the computational cost, we also propose a Low-rank Online Multi-modal DML (LOMDML) algorithm, which avoids the need of doing intensive positive semi-definite (PSD) projections and thus saves a significant amount of computational cost for DML on high-dimensional data. As a summary, the major contributions of this paper include:

- We present a novel framework of Online Multi-modal Distance Metric Learning (OMDML), which simultaneously learns optimal metrics on each individual modality and the optimal combination of the metrics from multiple modalities via efficient and scalable online learning;
- We further propose a low-rank OMDML algorithm which by significantly reducing computational costs for high-dimensional data without PSD projection;
- We offer theoretical analysis of the OMDML method;
- We conduct an extensive set of experiments to evaluate the performance of the proposed techniques for CBIR

Corresponding author: Steven C.H. HOI and Pengcheng WU are with the School of Information Systems, Singapore Management University, Singapore 178902, E-mail: {chhoi,pcwu}@smu.edu.sg
*Peilin Zhao is with the Data Analytics Department, Institute for Infocomm Research, A*STAR, Singapore 138632, E-mail: zhaop@i2r.a-star.edu.sg*
Chunyan Miao is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798, E-mail: ASCYMiao@ntu.edu.sg
Zhi-Yong Liu is with the State Key Lab of Management and Control for Complex System, Institute of Automation, Chinese Academy of Sciences, Beijing, China, E-mail: zhiyong.liu@ia.ac.cn

tasks using multiple types of features.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 first gives the problem formulation, and then presents our method of online multi-modal metric learning, followed by proposing an improved low-rank algorithm. Section 4 provides theoretical analysis for the proposed algorithms, Section 5 discusses our experimental results, and finally Section 6 concludes this work.

2 Related Work

Our work is related to three major groups of research: content-based image retrieval, distance metric learning, and online learning. In the following, we briefly review the closely related representative works in each group.

2.1 Content-based Image Retrieval

With the rapid growth of digital cameras and photo sharing websites, image retrieval has become one of the most important research topics in the past decades, among which content-based image retrieval is one of key challenging problems [1], [2], [3]. The objective of CBIR is to search images by analyzing the actual contents of the image as opposed to analyzing metadata like keywords, title and author, such that extensive efforts have been done for investigating various low-level feature descriptors for image representation [14]. For example, researchers have spent many years in studying various global features for image representation, such as color features [14], edge features [14], and texture features [15]. Recent years also witness the surge of research on local feature based representation, such as the bag-of-words models [16], [17] using local feature descriptors (e.g., SIFT [18]).

Conventional CBIR approaches usually choose rigid distance functions on some extracted low-level features for multimedia similarity search, such as the classical Euclidean distance or cosine similarity. However, there exists one key limitation that the fixed rigid similarity/distance function may not be always optimal because of the complexity of visual image representation and the main challenge of the semantic gap between the low-level visual features extracted by computers and high-level human perception and interpretation. Hence, recent years have witnesses a surge of active research efforts in design of various distance/similarity measures on some low-level features by exploiting machine learning techniques [19], [20], [21], among which some works focus on learning to hash for compact codes [22], [19], [23], [24], [25], and some others can be categorized into distance metric learning that will be introduced in the next subsection. Our work is also related to multimodal/multiview studies, which have been widely studied on image classification and object recognition fields [26], [27], [28], [29]. However, it is usually hard to exploit these techniques directly on CBIR because (i) in general, image classes will not be given explicitly on CBIR tasks, (ii) even if classes are given, the number will be very large, (iii) image datasets tend to be much larger on CBIR than on classification tasks. We thus exclude the direct comparisons to such existing works in this paper. There are still some other open issues in CBIR studies, such as the efficiency and scalability of the retrieval process that often requires an effective indexing scheme, which are out of this paper's scope.

2.2 Distance Metric Learning

Distance metric learning has been extensively studied in both machine learning and multimedia retrieval communities [30], [7], [31], [32], [33], [34], [35], [36]. The essential idea is to learn an optimal metric which minimizes the distance between similar/related images and simultaneously maximizes the distance between dissimilar/unrelated images. Existing DML studies can be grouped into different categories according to different learning settings and principles. For example, in terms of different types of constraint settings, DML techniques are typically categorized into two groups:

- Global supervised approaches [30], [7]: to learn a metric on a global setting, e.g., all constraints will be satisfied simultaneously;
- Local supervised approaches [32], [33]: to learn a metric in the local sense, e.g., the given local constraints from neighboring information will be satisfied.

Moreover, according to different training data forms, DML studies in machine learning typically learn metrics directly from explicit class labels [32], while DML studies in multimedia mainly learn metrics from side information, which usually can be obtained in the following two forms:

- Pairwise constraints [7], [9]: A must-link constraint set $\mathcal{S}$ and a cannot-link constraint set $\mathcal{D}$ are given, where a pair of images $(\mathbf{p}_i, \mathbf{p}_j) \in \mathcal{S}$ if $\mathbf{p}_i$ is related/similar to $\mathbf{p}_j$, otherwise $(\mathbf{p}_i, \mathbf{p}_j) \in \mathcal{D}$. Some literature uses the term equivalent/positive constraint in place of "must-link", and the term inequivalent/negative constraint in place of "cannot-link".
- Triple constraints [20]: A triplet set $\mathcal{P}$ is given, where $\mathcal{P} = \{(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-) | (\mathbf{p}_t, \mathbf{p}_t^+) \in \mathcal{S}; (\mathbf{p}_t, \mathbf{p}_t^-) \in \mathcal{D}, t = 1, \ldots, T\}$, $\mathcal{S}$ contains related pairs and $\mathcal{D}$ contains unrelated pairs, i.e., $\mathbf{p}$ is related/similar to $\mathbf{p}^+$ and $\mathbf{p}$ is unrelated/dissimilar to $\mathbf{p}^-$. T denotes the cardinality of entire triplet set.

When only explicit class labels are provided, one can also construct side information by simply considering relationships of instances in same class as related, and relationships of instances belonging to different classes as unrelated. In our works, we focus on triple constraints.

Finally, in terms of learning methodology, most existing DML studies generally employ batch learning methods which often assume the whole collection of training data must be given before the learning task and train a model from scratch, except for a few recent DML studies which begin to explore online learning techniques [37], [38]. All these works generally address single-modal DML, which is different from our focus on multi-modal DML. We also note that our work is very different from the existing multiview DML study [26] which is concerned with regular classification tasks by learning a metric on training data with explicit class labels, making it difficult to be compared with our method directly. We note that our work is different from another multimodal learning study in [39] which addresses a very different problem of search-based face

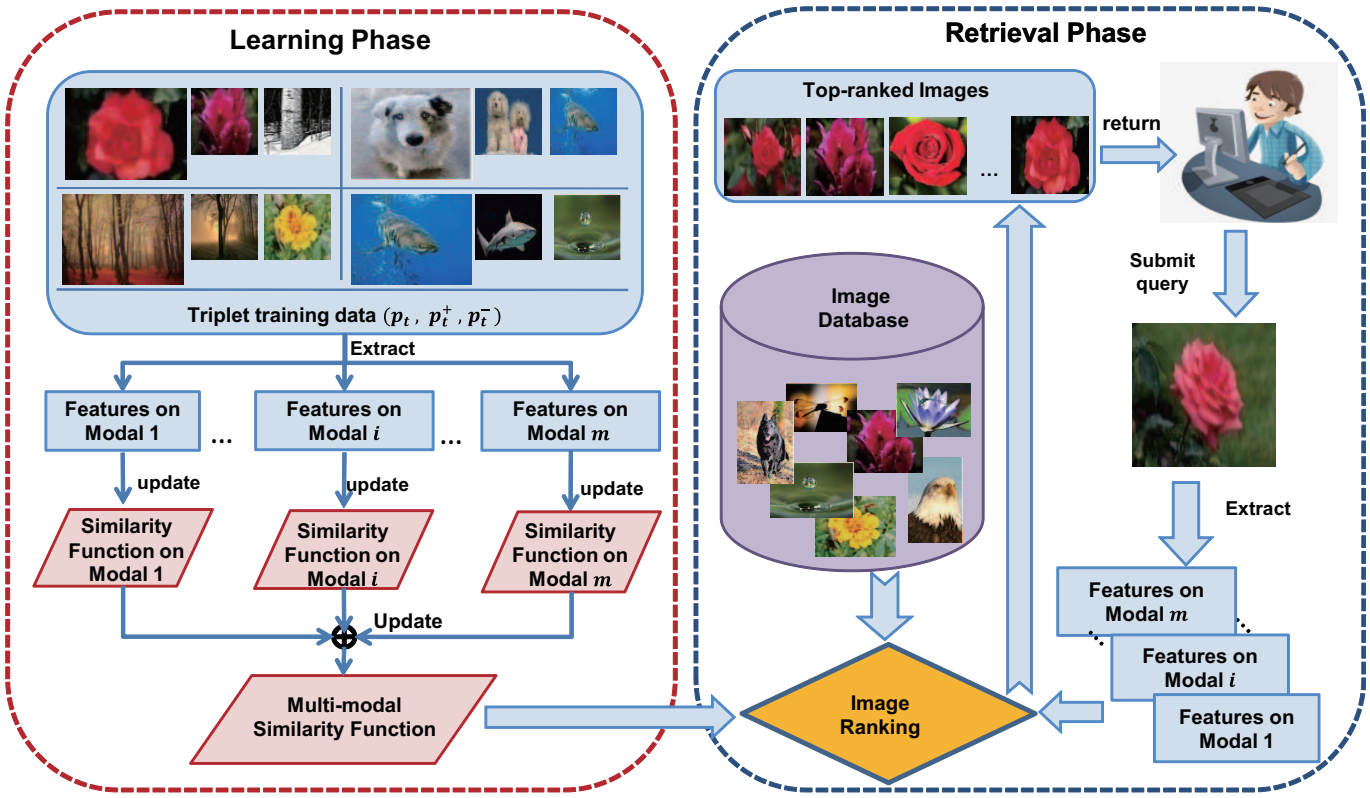


Fig. 1. Overview of the proposed multi-modal distance metric learning scheme for multi-modal retrieval in CBIR

annotation where their multimodal learning is formulated with a batch learning task for optimizing a specific loss function tailored for search-based face annotation tasks from weakly labeled data. Finally, we note that our work is also different from some existing distance learning studies that learn nonlinear distance functions using kernel or deep learning methods [21], [40], [35]. In comparison to the linear distance metric learning methods, kernel methods usually may achieve better learning accuracy in some scenarios, but falls short in being difficult to scale up for large-scale applications due to the curse of kernelization, i.e., the learning cost increases dramatically when the number of training instances increases. Thus, our empirical study is focused on direct comparisons to the family of linear methods.

2.3 Online Learning

Our work generally falls in the category of online learning methodology, which has been extensively studied in machine learning [41], [42]. Unlike batch learning methods that usually suffer from expensive re-training cost when new training data arrive, online learning sequentially makes a highly efficient (typically constant) update for each new training data, making it highly scalable for large-scale applications. In general, online learning operates on a sequence of data instances with time stamps. At each time step, an online learning algorithm processes an incoming example by first predicting its class label; after the prediction, it receives the true class label which is then used to measure the suffered loss between the predicted label and the true label; at the end of each time step, the model is updated with the loss whenever it is nonzero. The overall objective of an online learning task is to minimize the cumulative loss over the entire sequence of received instances.

In literature, a variety of algorithms have been proposed for online learning [43], [44], [45], [46], [47]. Some well-known examples include the Hedge algorithm for online prediction with expert advice [48], the Perceptron algorithm [43], the family of passive-Aggressive (PA) learning algorithms [44], and the online gradient descent algorithms [49]. There is also some study that attempts to improve the scalability of online kernel methods, such as [50] which proposed a bounded online gradient descent for addressing online kernel-based classification tasks. In this work, we apply online learning techniques, i.e., the Hedge, PA, and online gradient descent algorithms, to tackle the multi-modal distance metric learning task for content-based image retrieval. Besides, we note that this work was partially inspired by the recent study of online multiple kernel learning which aims to address online classification tasks using multiple kernels [51]. In the following, we give a brief overview of several popular online learning algorithms.

2.3.1 *Hedge Algorithms*

The Hedge algorithm [48], [52] is a learning algorithm which aims to dynamically combine multiple strategies in an optimal way, i.e., making the final cumulative loss asymptomatically approach that of the best strategy. Its key idea is to maintain a dynamic weigh-distribution over the set of strategies. During the online learning process, the distribution is updated according to the performance of those strategies. Specifically, the weight of every strategy is decreased exponentially with

respect to its suffered loss, making the overall strategy approaching the best strategy.

2.3.2 Passive-Aggressive Learning

As a classical well-known online learning technique, the Perceptron algorithm [43] simply updates the model by adding an incoming instance with a constant weight whenever it is misclassified. Recent years have witnessed a variety of algorithms proposed to improve Perceptron [53], [44], which usually follow the principle of maximum margin learning in order to maximize the margin of the classifier. Among them, one of the most notable approaches is the family of Passive-Aggressive (PA) learning algorithms [44], which updates the model whenever the classifier fails to produce a large margin on the incoming instance. In particular, the family of online PA learning is formulated to trade off the minimization of the distance between the target classifier and the previous classifier, and the minimization of the loss suffered by the target classier on the current instance. The PA algorithms enjoy good efficiency and scalability due to their simple closed-form solutions. Finally, both theoretical analysis and most empirical studies demonstrate the advantages of the PA algorithms over the classical Perceptron algorithm.

2.3.3 Online Gradient Descent

Besides Perceptron and PA methods, another well-known online learning method is the family of Online Gradient Descent (OGD) algorithms, which applies the family of online convex optimization techniques to optimize some particular objective function of an online learning task [49]. It enjoys solid theoretical foundation of online convex optimization, and thus works effectively in empirical applications. When the training data is abundant and computing resources are comparatively scarce, some existing studies showed that a properly designed OGD algorithm can asymptotically approach or even outperform a respective batch learning algorithm [54].

3 Online Multi-modal Distance Metric Learning

3.1 Overview

In literature, many techniques have been proposed to improve the performance of CBIR. Some existing studies have made efforts on investigating novel low-level feature descriptors in order to better represent visual content of images, while others have focused on the investigation of designing or learning effective distance/similarity measures based on some extracted low-level features. In practice, it is hard to find a single best low-level feature representation that consistently beats the others at all scenarios. Thus, it is highly desirable to explore machine learning techniques to automatically combine multiple types of diverse features and their respective distance measures. We refer to this open research problem as a multi-modal distance metric learning task, and present two new algorithms to solve it in this section. Figure 1 illustrates the system flow of the proposed multi-modal distance metric learning scheme for content-based image retrieval, which consists of two phases, i.e., learning phase and retrieval phase. The goal is to learn the distance metrics in the learning phase in order to facilitate the image ranking task in the retrieval phase. We note that these two phases may operate concurrently in practice, where the learning phase may never stop by learning from endless stream training data.

During the learning phase, we assume triplet training data instances arrive sequentially, which is natural for a real-world CBIR system. For example, in online relevance feedback, a user is often asked to provide feedback to indicate if a retrieved image is related or unrelated to a query; as a result, users' relevance feedback log data can be collected to generate the training data in a sequential manner for the learning task [55]. Once a triplet of images is received, we extract different low-level feature descriptors on multiple modalities from these images. After that, every distance function on a single modality can be updated by exploiting the corresponding features and label information. Simultaneously, we also learn the optimal combination of different modalities to obtain the final optimal distance function, which is applied to rank images in the retrieval phase.

During the retrieval phase, when the CBIR system receives a query from users, it first applies the similar approach to extract low-level feature descriptors on multiple modalities, then employs the learned optimal distance function to rank the images in the database, and finally presents the user with the list of corresponding top-ranked images. In the following, we first give the notation used throughout the rest of this paper, and then formulate the problem of multi-modal distance metric learning followed by presenting online algorithms to solve it.

3.2 Notation

For the notation used in this paper, we use bold upper case letter to denote a matrix, for example, $\mathbf{M} \in \mathbb{R}^{n \times n}$, and bold lower case letter to denote a vector, for example, $\mathbf{p} \in \mathbb{R}^n$. We adopt $\mathbf{I}$ to denote an identity matrix. Formally, we define the following terms and operates:

- m: the number of modalities (types of features).
- n_i: the dimensionality of the i-th visual feature space (modality).
- $\mathbf{p}^{(i)}$: the i-th type of visual feature (modality) of the corresponding image $\mathbf{p}^{(i)} \in \mathbb{R}^{n_i}$.
- $\mathbf{M}^{(i)}$: the optimal distance metric on the i-th modality, where $\mathbf{M}^{(i)} \in \mathbb{R}^{n_i \times n_i}$.
- $\mathbf{W}^{(i)}$: a linear transformation matrix by decomposing $\mathbf{M}^{(i)}$, such that, $\mathbf{M}^{(i)} = {\mathbf{W}^{(i)}}^T \mathbf{W}^{(i)}$, $\mathbf{W}_i \in \mathbb{R}^{r_i \times n_i}$, where r_i is the dimensionality of projected feature space.
- $\mathcal{S}$: a positive constraint set, where a pair $(\mathbf{p}_i, \mathbf{p}_j) \in \mathcal{S}$ if and only if $\mathbf{p}_i$ is related/similar to $\mathbf{p}_j$.
- $\mathcal{D}$: a negative constraint set, where a pair $(\mathbf{p}_i, \mathbf{p}_j) \in \mathcal{S}$ if and only if $\mathbf{p}_i$ is unrelated/dissimilar to $\mathbf{p}_j$.
- $\mathcal{P}$: a triplet set, where $\mathcal{P} = \{(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-) | (\mathbf{p}_t, \mathbf{p}_t^+) \in \mathcal{S}; (\mathbf{p}_t, \mathbf{p}_t^-) \in \mathcal{D}, t = 1, \ldots, T\}$, where T denotes the cardinality of entire triplet set.
- $d_i(\mathbf{p}_2, \mathbf{p}_2)$: the distance function of two images $\mathbf{p}_1$ and $\mathbf{p}_2$ on the i-th type of visual feature (modality).

When only one modality is considered, we will omit the superscript (i) or subscript i in the above terms.

3.3 Problem Formulation

Our goal is to learn a distance function from side information for content-based image retrieval. We restrict our discussion for learning the family of Mahalanobis distances. In particular, for any two images $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{R}^n$, where n is the dimensionality of represented feature space, we aim to learn an optimal distance metric $\mathbf{M}$ to calculate the distance between $\mathbf{p}_1$ and $\mathbf{p}_2$ as the following distance function:

$$d(\mathbf{p}_1, \mathbf{p}_2) = (\mathbf{p}_1 - \mathbf{p}_2)^\top \mathbf{M}(\mathbf{p}_1 - \mathbf{p}_2); \mathbf{M} \succeq 0, \quad (1)$$

where $\mathbf{M} \succeq 0$ denotes that $\mathbf{M}$ is a positive semi-definite (PSD) matrix, i.e., $\mathbf{p}^\top \mathbf{M}\mathbf{p} \geq 0$ for any nonzero real vector $\mathbf{p} \in \mathbb{R}^n$. Obviously, if one chooses $\mathbf{M}$ as the identity matrix $\mathbf{I}$, the above formula is reduced to the (square) Euclidean distance.

To formulate the learning task, we assume a collection of training data instances are given (sequentially) in the form of triplet constraints, i.e., $\mathcal{P} = \{(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-), t = 1, \ldots, T\}$, where each triplet indicates the relationship of three images, i.e., image $\mathbf{p}_t$ is similar to image $\mathbf{p}_t^+$ and dissimilar to $\mathbf{p}_t^-$. Typically, we can pose such a triplet relationship as the following constraint

$$d(\mathbf{p}_t, \mathbf{p}_t^+) \leq d(\mathbf{p}_t, \mathbf{p}_t^-) - 1; \forall t = 1, \ldots, T; \quad (2)$$

where -1 is a margin parameter to ensure a sufficiently large difference.

The above discussion generally assumes DML on single-modal data. We now generalize it to multi-modal data. In particular, we assume each image can be represented by a total of m feature spaces (modalities) and assume each feature space $\mathcal{F}_i$ is a n_i-dimensional vector space, i.e., $\mathcal{F}_i = \mathbb{R}^{n_i}$. The general idea of our multi-modal distance metric leaning is to learn a separate optimal distance metric $\mathbf{M}^{(i)} \in \mathbb{R}^{n_i \times n_i}$ for each feature space as

$$d_i(\mathbf{p}_1^{(i)}, \mathbf{p}_2^{(i)}) = (\mathbf{p}_1^{(i)} - \mathbf{p}_2^{(i)})^\top \mathbf{M}^{(i)}(\mathbf{p}_1^{(i)} - \mathbf{p}_2^{(i)}); \mathbf{M}^{(i)} \succeq 0,$$

and meanwhile learn an optimal combination of the distance functions from different modalities to obtain the final optimal distance function:

$$\begin{aligned} d(\mathbf{p}_1, \mathbf{p}_2) &= \sum_{i=1}^m \theta^{(i)} d_i(\mathbf{p}_1^{(i)}, \mathbf{p}_2^{(i)}) \\ &= \sum_{i=1}^m \theta^{(i)} (\mathbf{p}_1^{(i)} - \mathbf{p}_2^{(i)})^\top \mathbf{M}^{(i)} (\mathbf{p}_1^{(i)} - \mathbf{p}_2^{(i)}) \end{aligned}$$

where $\theta^{(i)} \in [0, 1]$ denotes the combination weight for the i-th modality and $\mathbf{p}_1^{(i)}, \mathbf{p}_2^{(i)} \in \mathcal{F}_i$ denote the visual features on the space of i-th modality. In the following, without loss of clarity, we will simplify denote $d_i(\mathbf{p}_1^{(i)}, \mathbf{p}_2^{(i)})$ as $d_i(\mathbf{p}_1, \mathbf{p}_2)$ by removing the superscript.

To simultaneously learn both the optimal combination weights $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(m)})$ and the optimal individual distance metric $\{\mathbf{M}^{(i)} | i = 1 \ldots, m\}$, we cast the multi-modal distance metric learning problem into the following optimization task:

$$\min_{\boldsymbol{\theta} \in \Delta} \min_{\mathbf{M}^{(i)} \succeq 0} \frac{1}{2} \sum_{i=1}^m \|\mathbf{M}^{(i)}\|_F^2 + C \sum_{t=1}^T \ell_t\big((\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-); d\big) \quad (3)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\Delta = \{\boldsymbol{\theta} | \sum_{i=1}^m \theta^{(i)} = 1, \theta^{(i)} \in [0, 1], \forall i\}$ and $\ell_t(\cdot)$ is a loss function such as

$$\ell((\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-); d) = \max(0, d(\mathbf{p}_t, \mathbf{p}_t^+) - d(\mathbf{p}_t, \mathbf{p}_t^-) + 1).$$

The constraints in Eqn.(2) are implicitly imposed in the above hinge loss function, and C is a regularization parameter to prevent overfitting.

3.4 OMDML Algorithm

One way is to directly solve the optimization task in Eqn.(3) via a batch learning approach. This is however not a good solution primarily for two key reasons:

- A critical drawback of such a batch training solution is that it suffers from extremely high re-training cost, i.e, whenever there is a new training instance, the entire model has to be completely re-trained from scratch, making it non-scalable for real-world applications;
- Beside, solving Eqn.(3) directly can be computationally very expensive for a large amount of training data;

To address these challenges, we present an online learning algorithm to tackle the multi-modal distance metric learning task.

Algorithm 1 OMDML — Online Multi-modal DML

INPUT:
- Discount weight: $\beta \in (0, 1)$
- regularization parameter: $C > 0$
- margin parameter: $\gamma \geq 0$

Initialization:
- $\theta_1^{(i)} = 1/m$, $\forall i = 1, \ldots, m$
- $\mathbf{M}_{b1}^{(i)} = \mathbb{I}$, $\forall i = 1, \ldots, m$

for $t = 1, 2, \ldots, T$ **do**
  Receive: $(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-)$
  $f_t^{(i)} = d_i(\mathbf{p}_t, \mathbf{p}_t^+) - d_i(\mathbf{p}_t, \mathbf{p}_t^-), \forall i = 1, \ldots, m$
  $f_t = \sum_{i=1}^m \theta_t^{(i)} f_t^{(i)}$
  if $f_t + \gamma > 0$ **then**
    for $i = 1, 2, \ldots, m$ **do**
      Set $z_t^{(i)} = \mathbb{I}(f_t^{(i)} > 0)$
      Update $\theta_{t+1}^{(i)} \leftarrow \theta_t^{(i)} \beta^{z_t^{(i)}}$
      Update $\mathbf{M}_{t+1}^{(i)} \leftarrow \mathbf{M}_t^{(i)} - \tau_t^{(i)} \mathbf{V}_t^{(i)}$ by Eq. (5)
      Update $\mathbf{M}_{t+1}^{(i)} \leftarrow PSD(\mathbf{M}_{t+1}^{(i)})$
    end for
    $\Theta_{t+1} = \sum_{i=1}^m \theta_{t+1}^{(i)}$
    $\theta_{t+1}^{(i)} \leftarrow \theta_{t+1}^{(i)} / \Theta_{t+1}, \forall i = 1, \ldots, m$
  end if
end for

The key challenge to online multi-modal distance metric learning tasks is to develop an efficient and scalable learning scheme that can optimize both the distance metric on each individual modality and meanwhile optimize the combinational weights of different modalities. To this end, we propose to explore an online distance metric learning algorithm, i.e., a variant of OASIS [20] and PA [44], to learn the individual distance metric, and apply the well-known Hedge algorithm [48]

to learn the optimal combinational weights. We discuss each of the two learning tasks in detail below.

Let us denote by $\mathbf{M}_t^{(i)}$ the matrix on the i-th modality at step t. To learn the optimal metric $\mathbf{M}_t^{(i)}$ on an individual modality, following the similar ideas of OASIS [20] and PA [44], we can formulate the optimization task of the online distance metric learning as follows:

$$\begin{aligned} \mathbf{M}_{t+1}^{(i)} = \arg\min_{\mathbf{M}} \quad & \frac{1}{2}\|\mathbf{M} - \mathbf{M}_t^{(i)}\|_F + C\xi, \\ s.t. \quad & \ell((\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-); d_i) \le \xi, \quad \xi \ge 0 \end{aligned} \tag{4}$$

It is not difficult to derive the closed-form solution:

$$\mathbf{M}_{t+1}^{(i)} = \mathbf{M}_t^{(i)} - \tau_t^{(i)} \mathbf{V}_t^{(i)} \tag{5}$$

where $\tau_t^{(i)}$ and $\mathbf{V}_t^{(i)}$ are computed as follows:

$$\begin{aligned} \tau_t^{(i)} &= \min(C, \ell((\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-); d_i)/\|\mathbf{V}_t^{(i)}\|_F^2), \\ \mathbf{V}_t^{(i)} &= (\mathbf{p}_t - \mathbf{p}_t^+)(\mathbf{p}_t - \mathbf{p}_t^+)^\top - (\mathbf{p}_t - \mathbf{p}_t^-)(\mathbf{p}_t - \mathbf{p}_t^-)^\top. \end{aligned}$$

In the above, we omit the superscript (i) for each $\mathbf{p}_t$.

One main issue of the above solution, as existed in OASIS [20], is that it does not guarantee the resulting matrix $\mathbf{M}_{t+1}^{(i)}$ is positive semi-definite (PSD), which is not desirable for DML. To fix this issue, at the end of each learning iteration, we will need to perform a PSD projection of the matrix $\mathbf{M}$ onto the PSD domain:

$$\mathbf{M}_{t+1}^{(i)} \leftarrow PSD(\mathbf{M}_{t+1}^{(i)}).$$

Another key task of multi-modal DML is to learn the optimal combinational weights $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(m)})$, where $\theta^{(i)}$ is set to $1/m$ at the beginning of the learning task. We apply the well-known Hedge algorithm [48] to update the combinational weights online, which is a simple and effective algorithm for online learning with expert advice. In particular, given a triplet training instance $(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-)$, at the end of each online learning iteration, the weight is updated as follows:

$$\theta_{t+1}^{(i)} = \frac{\theta_t^{(i)} \beta^{z_t^{(i)}}}{\sum_{i=1}^m \theta_t^{(i)} \beta^{z_t^{(i)}}} \tag{6}$$

where $\beta \in (0, 1)$ is a discounting parameter to penalize the poor modality, and $z_t^{(i)}$ is an indicator of ranking result on the current instance, i.e., $z_t^{(i)} = \mathbb{I}(f_t^{(i)} > 0) = \mathbb{I}(d_i(\mathbf{p}_t, \mathbf{p}_t^+) - d_i(\mathbf{p}_t, \mathbf{p}_t^-) > 0)$ which outputs 1 when $f_t^{(i)} = d_i(\mathbf{p}_t, \mathbf{p}_t^+) - d_i(\mathbf{p}_t, \mathbf{p}_t^-) > 0$ and 0 otherwise. In particular, $f_t^{(i)} > 0$, namely $d_i(\mathbf{p}_t, \mathbf{p}_t^+) > d_i(\mathbf{p}_t, \mathbf{p}_t^-)$, indicates the current i-th metric makes a mistake on predicting the ranking of the triplet $(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-)$.

Finally, Algorithm 1 summarizes the details of the proposed Online Multi-modal Distance Metric Learning (OMDML) algorithm.

Remark on Space and Time complexity. The space complexity of the algorithm is $\mathcal{O}(\sum_{i=1}^m {n_i}^2)$. Denoting $n = \max(n_1, \ldots, n_m)$, the worst-case space complexity is simply $\mathcal{O}(m \times n^2)$. The overall time complexity is linear with respect to T — the total number of training triplets. The most computationally intensive step is the PSD projection step, which can be $\mathcal{O}(n^3)$ for a dense matrix. Hence, the worst-case time overall complexity is $\mathcal{O}(T \times m \times n^3)$.

3.5 Low-Rank Online Multi-modal Distance Metric Learning Algorithm

One critical drawback of the proposed OMDML algorithm in Algorithm 1 is the PSD projection step, which can be computationally intensive when some feature space is of high dimensionality. In this section, we present a low-rank learning algorithm to significantly improve the efficiency and scalability of OMDML.

Instead of learning a full-rank matrix, for each $\mathbf{M}^{(i)}$, our goal is to learn a low-rank decomposition, i.e.,

$$\mathbf{M}^{(i)} := \mathbf{W}^{(i)\top} \mathbf{W}^{(i)},$$

where $\mathbf{W}_i \in \mathbb{R}^{r_i \times n_i}$ and $r_i \ll n_i$. Thus, for any two images $\mathbf{p}_1$ and $\mathbf{p}_2$, the distance function on the i-th modality can be expressed as:

$$d_i(\mathbf{p}_1, \mathbf{p}_2) = (\mathbf{p}_1 - \mathbf{p}_2)^T \mathbf{W}^{(i)\top} \mathbf{W}^{(i)} (\mathbf{p}_1 - \mathbf{p}_2)$$

Following the similar idea in the previous section, we can apply online learning techniques to solve $\mathbf{W}_t^{(i)}$ and $\boldsymbol{\theta}_t$, respectively. In this section, we consider the Online Gradient Descent (OGD) approach to solve $W_t^{(i)}$. In particular, we denote by

$$\begin{aligned} \ell_t^{(i)} &= \ell((\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-); d_i) \\ &= \max(0, d(\mathbf{p}_t, \mathbf{p}_t^+) - d(\mathbf{p}_t, \mathbf{p}_t^-) + 1), \end{aligned}$$

and introduce the following notation

$$\mathbf{q}_t = \mathbf{W}_t^{(i)} \mathbf{p}_t, \ \mathbf{q}_t^+ = \mathbf{W}_t^{(i)} \mathbf{p}_t^+, \ \mathbf{q}_t^- = \mathbf{W}_t^{(i)} \mathbf{p}_t^-,$$

we can compute the gradient of $\ell_t^{(i)}$ with respect to $\mathbf{W}^{(i)}$:

$$\begin{aligned} &\nabla_t \mathbf{W}^{(i)} = \frac{\partial \ell_t^{(i)}}{\partial \mathbf{W}^{(i)}} \\ &= \sum_{j=1}^{r_i} \left(\frac{\partial \ell_t^{(i)}}{\partial q_{j,t}} \frac{\partial q_{j,t}}{\partial \mathbf{W}^{(i)}} + \frac{\partial \ell_t^{(i)}}{\partial q_{j,t}^+} \frac{\partial q_{j,t}^+}{\partial \mathbf{W}^{(i)}} + \frac{\partial \ell_t^{(i)}}{\partial q_{j,t}^-} \frac{\partial q_{j,t}^-}{\partial \mathbf{W}^{(i)}} \right) \Big|_{\mathbf{W}^{(i)} = \mathbf{W}_t^{(i)}} \\ &= 2(-\mathbf{q}_t^+ + \mathbf{q}_t^-)\mathbf{p}_t^\top + 2(-\mathbf{q}_t + \mathbf{q}_t^+){\mathbf{p}_t^+}^\top + 2(\mathbf{q}_t - \mathbf{q}_t^-){\mathbf{p}_t^-}^\top, \end{aligned}$$

where $q_{j,t}$ is the j-th entry of $\mathbf{q}_t$.

We then follow the idea of Online Gradient Descent [49] to update $\mathbf{W}_{t+1}^{(i)}$ of each modality as follows:

$$\mathbf{W}_{t+1}^{(i)} \leftarrow \mathbf{W}_t^{(i)} - \eta \nabla_t \mathbf{W}^{(i)} \tag{7}$$

where η is a learning rate parameter.

Similarly, we also apply the Hedge algorithm as introduced in the previous section to update the combinational weight $\boldsymbol{\theta}_t$. Finally, Algorithm 2 summarizes the details of the proposed Low-rank Online Multi-modal Metric Learning algorithm (LOMDML).

Clearly this algorithm naturally preserves the PSD property of the resulting distance metric $\mathbf{M}^{(i)} = \mathbf{W}^{(i)\top} \mathbf{W}^{(i)}$ and thus avoids the needs of performing the intensive PSD projection. By assuming all $r_1 = \ldots = r_m = r$ and $n = \max(n_1, \ldots, n_m)$, the overall time complexity of the algorithm is $\mathcal{O}(T \times m \times r \times n)$.

Algorithm 2 LOMDML—Low-rank OMDML algorithm

INPUT:
- Discount weight parameter: $\beta \in (0,1)$
- Margin parameter: $\gamma > 0$
- Learning rate parameter: $\eta > 0$

Initialization: $\theta_1^{(i)} = 1/m$, $\mathbf{W}_t^{(i)}$, $\forall i = 1, \ldots, m$
for $t = 1, 2, \ldots, T$ **do**
  Receive: $(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-)$
  Compute: $f_t^{(i)} = d_i(\mathbf{p}_t, \mathbf{p}_t^+) - d_i(\mathbf{p}_t, \mathbf{p}_t^-), i = 1, \ldots, m$
  Compute: $f_t = \sum_{i=1}^m \theta_t^{(i)} f_t^{(i)}$
  if $f_t + \gamma > 0$ **then**
    for $i = 1, 2, \ldots, m$ **do**
      Set $z_t^{(i)} = \mathbb{I}(f_t^{(i)} > 0)$
      Update $\theta_{t+1}^{(i)} \leftarrow \theta_t^{(i)} \beta^{z_t^{(i)}}$
      $\mathbf{W}_{t+1}^{(i)} \leftarrow \mathbf{W}_t^{(i)} - \eta \nabla_t \mathbf{W}^{(i)}$ by Eq. (7)
    end for
    $\Theta_{t+1} = \sum_{i=1}^m \theta_{t+1}^{(i)}$
    $\theta_{t+1}^{(i)} \leftarrow \theta_{t+1}^{(i)} / \Theta_{t+1}$, $i = 1, \ldots, m$
  end if
end for

4 Theoretical Analysis

We now analyze the theoretical performance of the proposed algorithms. To be concise, we give a theorem for the bound of mistakes made by Algorithm 1 for predicting the relative similarity of the sequence of triplet training instances. The similar result can be derived for Algorithm 2.

For the convenience of discussions in this section, we define:

$$z_t^{(i)} = \mathbb{I}\left(f_t^{(i)} > 0\right),$$

where $\mathbb{I}(x)$ is an indicator function that outputs 1 when x is true and 0 otherwise. We further define the optimal margin similarity function error for $\mathbf{M}^{(i)}$ with respect to a collection of training examples $\mathcal{P} = \{(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-), t = 1, \ldots, T\}$ as

$$F(\mathbf{M}^{(i)}, \ell, \mathcal{P}) = \min_{\mathbf{M}^{(i)}} \left\{ \frac{\left[\|\mathbf{M}^{(i)} - \mathbf{I}\|_F^2 + 2C \sum_{t=1}^T \ell_t(d_i)\right]}{\min(C, 1)} \right\}$$

where $\ell_t(d_i)$ denotes $\ell((\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-); d_i)$. We then have the following theorem for the mistake bound of the proposed OMDML algorithm.

Theorem 1. *After receiving a sequence of T training examples, denoted by $\mathcal{P} = \{(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-), t = 1, \ldots, T\}$, the number of mistakes $\mathcal{M}$ on predicting the ranking of $(\mathbf{p}_t, \mathbf{p}_t^+, \mathbf{p}_t^-)$ made by running Algorithm 1, denoted by*

$$\mathcal{M} = \sum_{t=1}^T \mathbb{I}(f_t > 0) = \sum_{t=1}^T \mathbb{I}\left(\sum_{i=1}^m \theta_t^{(i)} f_t^{(i)} > 0\right)$$

is bounded as follows

$$\begin{aligned} \mathcal{M} &\leq \frac{2\ln(1/\beta)}{1-\beta} \min_{1 \leq i \leq m} \sum_{t=1}^T z_t^{(i)} + \frac{2\ln m}{1-\beta} \\ &\leq \frac{2\ln(1/\beta)}{1-\beta} \min_{1 \leq i \leq m} F(\mathbf{M}^{(i)}, \ell, \mathcal{P}) + \frac{2\ln m}{1-\beta} \end{aligned}$$

By choosing $\beta = \frac{\sqrt{T}}{\sqrt{T} + \sqrt{\ln m}}$, we then have

$$\mathcal{M} \leq 2\left(\left(1 + \sqrt{\frac{\ln m}{T}}\right) \min_{1 \leq i \leq m} F(\mathbf{M}^{(i)}, \ell, \mathcal{P}) + \ln m + \sqrt{T \ln m}\right)$$

In general, it is not difficult to prove the above theorem by combining the results of the Hedge algorithm and the PA online learning, similar to the technique used in [51]. More details about the proof can be found in the online supplemental file [1]. Basically the above theorem indicates that the total number of mistakes of the proposed algorithm is bounded by $O(\sqrt{T})$ compared with the optimal single metric.

5 Experiments

In this section, we conduct an extensive set of experiments to evaluate the efficacy of the proposed algorithms for similarity search with multiple types of visual features in CBIR.

5.1 Experimental Testbeds

We adopt four publicly-available image data sets in our experiments, which have been widely adopted for the benchmarks of content-based image retrieval, image classification and recognition tasks. TABLE 1 summarizes the statistics of these databases.

TABLE 1
List of image databases in our testbed.

Datasets	size	classes #	avg # per class
Caltech101	8,677	101	85.91
Indoor	15,620	67	233.14
ImageCLEF	7,157	20	367.85
Corel	5,000	50	100
ImageCLEFFlickr	1,007,157	21	47959.86

The first testbed is the "caltech101"[2], which has been widely adopted for object recognition and image retrieval [56], [20]. This dataset contains 101 object categories and 8,677 images.

The second testbed is the "indoor" dataset[3], which was used for recognizing indoor scenes [57]. This dataset consists of 67 indoor categories, and 15,620 images. The numbers of images in different categories are diverse, but each category contains at least 100 images. It is further divided into 5 subsets: store, home, public spaces, leisure, and working place. We simply consider it as a dataset of 67 categories and evaluate different algorithms on the whole indoor collection.

The third testbed is the "ImageCLEF" dataset[4], which was also used in [58]. It is a medical image dataset and has 7,157 images in 20 categories.

The fourth testbed is the "Corel" dataset [7], which consists of photos from COREL image CDs. It has 50 categories, each of which has exactly 100 images randomly selected from related examples in COREL image CDs.

1. http://omdml.stevenhoi.org/
2. http://www.vision.caltech.edu/Image_Datasets/Caltech101/
3. http://web.mit.edu/torralba/www/indoor.html
4. http://imageclef.org/

We also combine "ImageCLEF" with a collection of one million social photos crawled from Flickr, this larger set is named "ImageCLEFFlickr". We treat the Flickr photos as a special class of background noisy photos, which are mainly used to test the scalability of our algorithms.

5.2 Experimental Setup

For each database, we split the whole dataset into three disjoint partitions: a training set, a test set, and a validation set. In particular, we randomly choose 500 images to form a test set, and other 500 images to build up a validation set. The remaining images are used to form a training set for learning similarity functions.

To generate side information in the form of triplet instances for learning the ranking functions, we sample triplet constraints from the images in the training set according to their ground truth labels. Specifically, we generate a triplet instance by randomly sampling two images belonging to the same class and one image from a different class. In total, we generate 100K triplet instances for each standard dataset (except for the small-scale and large-scale experiments).

To fairly evaluate different algorithms, we choose their parameters by following the same cross validation scheme. For simplicity, we empirically set $r_i = r = 50$ for the i-th modality in the LOMDML algorithm and set the maximum iteration to 500 for LMNN. To evaluate the retrieval performance, we adopt the mean Average Precision (mAP) and top-K retrieval accuracy. As a widely used IR metric, mAP value averages the Average Precision (AP) value of all the queries, each of which denotes the area under precision-recall curve for a query. The precision value is the ratio of related examples over total retrieved examples, while the recall value is the ratio of related examples retrieved over total related examples in the database.

Finally, we run all the experiments on a Linux machine with 2.33GHz 8-core Intel Xeon CPU and 16GB RAM.

5.3 Diverse Visual Features for Image Descriptors

We adopt both global and local feature descriptors to extract features for representing images in our experiments. Each feature will correspond to one modality in the algorithm. Before the feature extraction, we have preprocessed the images by resizing all the images to the scale of 500×500 pixels while keeping the aspect ratio unchanged.

Specifically, for global features, we extract five types of features to represent an image, namely

- Color histogram and color moments ($n = 81$),
- Edge direction histogram ($n = 37$),
- Gabor wavelets transformation ($n = 120$),
- Local binary pattern ($n = 59$),
- GIST features ($n = 512$).

For local features, we extract the bag-of-visual-words representation using two kinds of descriptors:

- SIFT — we adopt the Hessian-Affine interest region detector with a threshold of 500;
- SURF — we use the SURF detector with a threshold of 500.

For the clustering step, we adopt a forest of 16 kd-trees and search 2048 neighbors to speed up the clustering task. By combining different descriptors (SIFT/SURF) and vocabulary sizes (200/1000), we extract four types of local features: SIFT200, SIFT1000, SURF200 and SURF1000. Finally, we adopt the TF-IDF weighing scheme to generate the final bag-of-visual-words for describing the local features. For all learning algorithms, we normalize the feature vectors to ensure that every feature entry is in $[0, 1]$.

5.4 Comparison Algorithms

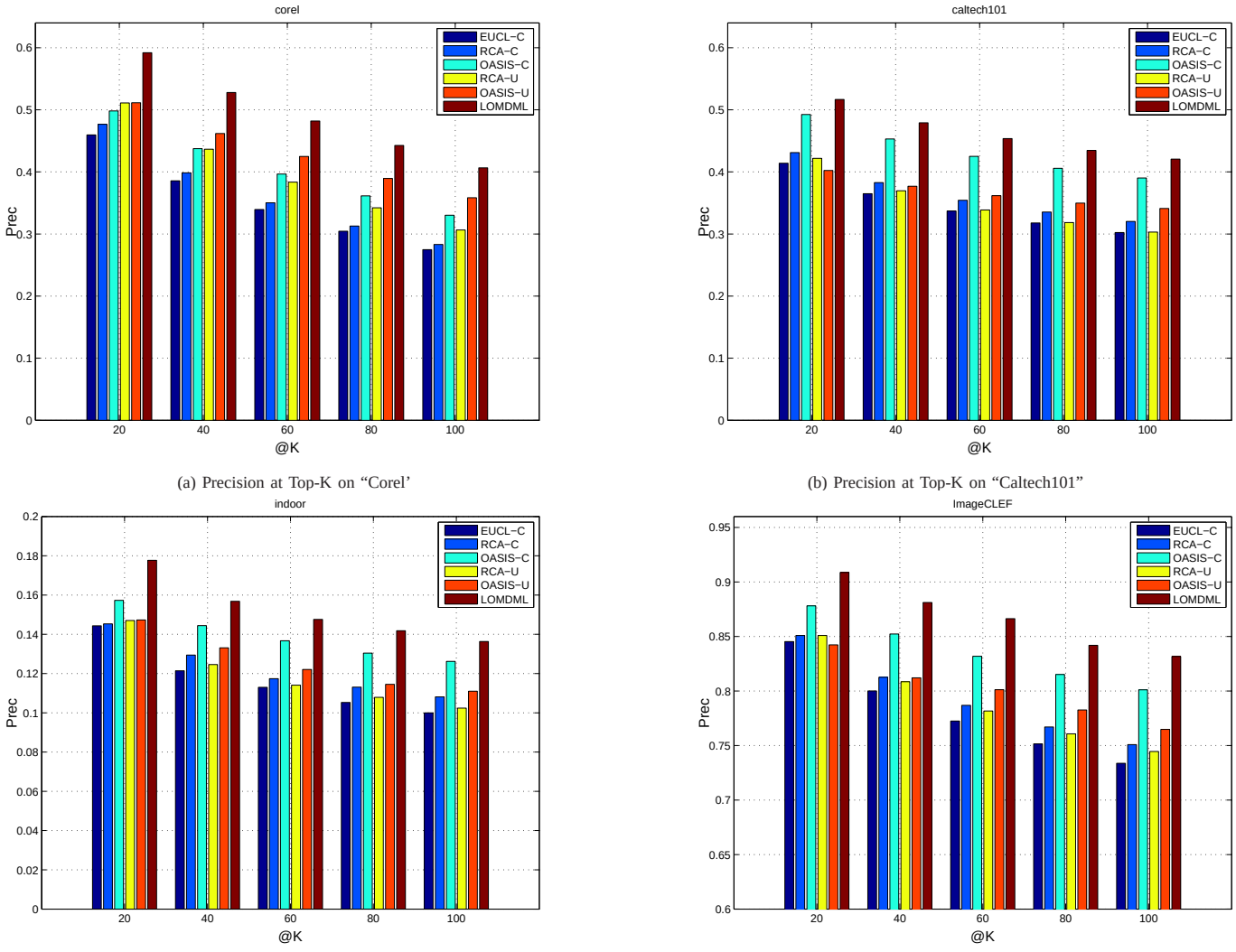
To extensively evaluate the efficacy of our algorithms, we compare the proposed two online multi-modal DML algorithms, i.e., OMDML and LOMDML, against a number of existing representative DML algorithms, including RCA [30], LMNN [32], and OASIS [20]. As a heuristic baseline method, we also evaluate the square Euclidean distance, denoted as "EUCL-*".

To adapt the existing DML methods for multi-modal image retrieval, we have implemented several variants of each DML algorithm by exploring three fusion strategies [59], [60]:

1) "**B**est" — applying DML for each modality individually and then selecting the best modality. We name these algorithms with suffix "-B", e.g., **RCA-B**, in which we first learn metrics over each modality separately on the training set by Relevance Component Analysis (RCA) [30]. After that, we validate the retrieval performance of all metrics on corresponding modality against the validation set, and then choose the modality with the highest mAP as the best modality. We report the mAP score over the best modality by ranking on test set with RCA.
2) "**C**oncatenation" — an early fusion approach by concatenating features of all modalities before applying DML. We name these algorithms with suffix "-C", e.g., **LMNN-C**, in which we first *concatenate* all types of features together, and then learn the optimal metric on this combined feature space by LMNN [32], and finally evaluate the mAP score on the optimal metric.
3) "**U**niform combination" — a late fusion approach by uniformly combining all modalities after metric learning. We name these algorithms with suffix "-U", e.g., **OASIS-U**, in which we first learn an optimal metric by OASIS [20] for each modality, and then *uniformly combine* all distance functions for the final ranking.

5.5 Evaluation on Small-Scale Datasets

In this section, we build four small-scale data sets, named "Caltech101(S)", "Indoor(S)", "COREL(S)" and "ImageCLEF(S)", from the corresponding standard datasets by first choosing 10 object categories, and then randomly sampling 50 examples from each category. We adopt 5 global features described above as the multi-modal inputs. To construct triplet constraints for online learning approaches, we generate all positive pairs (two images belong to the same class), and for each positive pair we randomly select an image from the other different classes to form a triplet. In total, about 10K triplets are generated for each dataset. TABLE 2 summarizes



(a) Precision at Top-K on "Corel'

(b) Precision at Top-K on "Caltech101"

(c) Precision at Top-K on "Indoor"

(d) Precision at Top-K on "ImageCLEF"

Fig. 2. Evaluation of average precision at Top-K results on the datasets

the evaluation results on the small-scale data sets, from which we can draw the following observations.

TABLE 2
Evaluation of the mAP performance.

Alg.	COREL(S)	Caltech101(S)	Indoor(S)	ImageCLEF(S)
Eucl-B	0.4431	0.4299	0.1726	0.4325
RCA-B	0.5097	0.4984	0.1915	0.4492
LMNN-B	0.4876	0.5462	0.1852	0.5231
OASIS-B	0.4445	0.5072	0.1884	0.4424
Eucl-C	0.5220	0.4306	0.1842	0.4431
RCA-C	0.6437	0.6156	0.2078	0.5927
LMNN-C	0.5816	0.5894	0.2027	0.5821
OASIS-C	0.5657	0.5441	0.2017	0.5618
Eucl-U	0.5220	0.4306	0.1842	0.4431
RCA-U	0.5625	0.4860	0.1894	0.4909
LMNN-U	0.6026	0.4282	0.2007	0.4647
OASIS-U	0.5679	0.5419	0.1989	0.5338
OMDML	0.6620	0.6543	0.2113	0.6824
LOMDML	**0.6975**	**0.6646**	**0.2250**	**0.7080**

First of all, the two kinds of fusion strategies, i.e., early fusion (with suffix"-C") and late fusion (with suffix"-U"), generally tend to perform better than the best single metric approaches (with suffix"-B"). This is primarily because combining multiple types of features with learning could better explore the potential of all the features, which validates the importance of the proposed technique.

Second, some of the uniformly combination algorithms (i.e., the late fusion strategy) failed to outperform the best single metric approach in some cases, e.g., "RCA-U" (compared with "RCA-B") and "LMNN-U" (compared with "LMNN-B") on Caltech101(S). This implies that uniform concatenation is not optimal to combine different kinds of features. Thus, it is critical to identify the effective features via machine learning and then assign them higher weights.

Third, among all the compared algorithms, the proposed OMDML and LOMDML algorithms outperform the other algorithms. Finally, it is interesting to observe that the proposed low-rank algorithm (LOMDML) not only improves the efficiency and scalability of OMDML, but also enhances the retrieval accuracy. This is probably because by learning metrics in intrinsic lower-dimensional space, we may potentially avoid the impact of overfitting and noise issues.

TABLE 3
Running time cost (in sec.) on "COREL(S)".

RCA-C	LMNN-C	OASIS-C	RCA-U
5.07	1442.66	404.35	2.91
LMNN-U	OASIS-U	OMDML	LOMDML
858.94	376.77	34765.13	22.11

TABLE 3 shows the running CPU time cost (in seconds) on the "COREL(S)" data set. We can see that, the running time of LOMDML results in a speedup factor of 10 comparison to OASIS, and the gain in efficiency will increase when the data set gets larger or the data dimensionality increases. Conversely, OMDML has the extremely high computational cost because a PSD projection is performed after each iteration, which can be $O(n^3)$ for a dense matrix. A possible solution to tackle this problem is that in we could perform the PSD projection after a bunch of iterations, instead of after each iteration.

5.6 Evaluation on the Standard Datasets

TABLE 4
Evaluation of the mAP performance.

Alg.	COREL	Caltech101	Indoor	ImageCLEF
Eucl-B	0.1877	0.2187	0.0469	0.5523
RCA-B	0.2305	0.2837	0.0499	0.6010
OASIS-B	0.1958	0.3025	0.0522	0.6723
Eucl-C	0.2628	0.2259	0.0559	0.5752
RCA-C	0.2714	0.2473	0.0604	0.6272
OASIS-C	0.3202	0.3660	0.0726	0.7394
Eucl-U	0.2628	0.2259	0.0559	0.5752
RCA-U	0.2992	0.2413	0.0565	0.6161
OASIS-U	0.3594	0.3243	0.0705	0.6891
LOMDML	**0.4137**	**0.4128**	**0.0804**	**0.8155**

We further evaluate the proposed algorithms on standard-sized image datasets. We exclude LMNN and OMDML because of their extremely high computational cost. Following the standard experimental setup with 5 global features and 4 local features, TABLE 4 summarizes the experimental results, Figure 2 presents the top-K precisions on four datasets and TABLE 5 shows the running time cost on the COREL dataset with 100K triplet instances. From the results, we observed that the proposed LOMDML algorithm considerably surpasses all the other approaches for most cases. This clearly validates the efficacy of the proposed algorithm for learning effective metrics on multi-modal data. Finally, in terms of the time cost, the proposed LOMDML algorithm is considerably more efficient and scalable than the other algorithms, making it practical for large-scale applications.

TABLE 5
Running time (in sec.) on "COREL".

RCA-C	OASIS-C	RCA-U	OASIS-U	LOMDML
468.19	65060.93	184.3	8781.54	789.81

Remark. We note that the learnt metric/function can be easily integrated into a generic image indexing and retrieval system, i.e., performing a linear projection for each image instance $\mathbf{p}$ by $\mathbf{p} \leftarrow \mathbf{W}\mathbf{p}$. The time cost for retrieval on OMDML is thus the same as the original Euclidean distance, while the time cost on LOMDML is the same as Euclidean distance on dimension-reduced feature space. To avoid the trivial redundant results, we thus skip the time cost evaluation of retrieval in our experiments.

5.7 Evaluation of online mistake rate of individual metric learning on each single modality

To further examine how the proposed LOMDML algorithm performs in comparison to individual metric learning on each single modality, we evaluate the online average mistake rate of the proposed LOMDML algorithm and single-modal metric learning schemes on each individual modality. Figure 3 shows the experimental results on the "COREL" data set. Several observations can be drawn from the results as follows.

First of all, we notice that for all the schemes, the online cumulative mistake rate consistently decreases when the number of iterations increases in the online learning process. Second, among all kinds of features, we found that the scheme of single-modal metric learning on "Surf1000" achieved the best performance. Finally, by comparing the proposed LOMDML scheme and the best single-modal metric learning, we found that LOMDML consistently achieves the smaller mistake rate than that of the best single-modal metric learning scheme in the entire online learning process. This encouraging result again validates the efficacy of the proposed multi-modal online learning scheme for combining multiple modalities in an effective way.

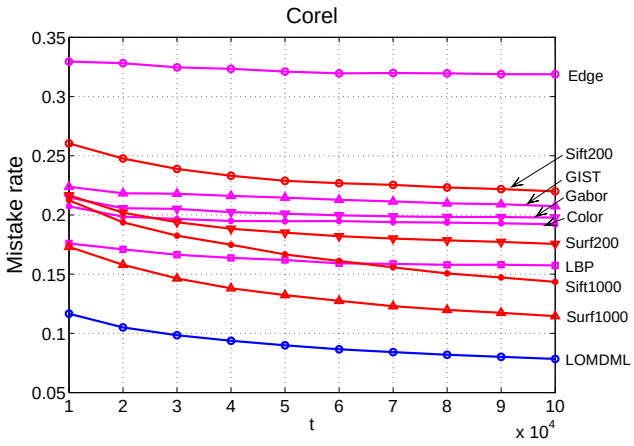


Fig. 3. Evaluation of online mistake rates of LOMDML and single-modal metric learning on individual modalities on the "Corel" dataset

5.8 Comparison with Online Multi-modal Distance Learning (OMDL) with Multiple Kernels

In this section, we compare the proposed LOMDML algorithm with an existing Online Multi-modal Distance Learning method (OMDL-LR) [40], which is a kernel-based low-rank online learning approach to learning distance functions on multi-modal data by combining multiple kernels. We evaluate

TABLE 6
Comparison between LOMDML and OMDL-LR (gaussianmeanvar).

Metric	Dataset	LOMDML	OMDL-LR
mAP	COREL(S)	0.6975	0.6693
	Caltech101(S)	0.6646	0.5994
	Indoor(S)	0.2250	0.2088
	ImageCLEF(S)	0.7080	0.6729
Time cost (in sec.)	COREL(S)	22.11	209.57

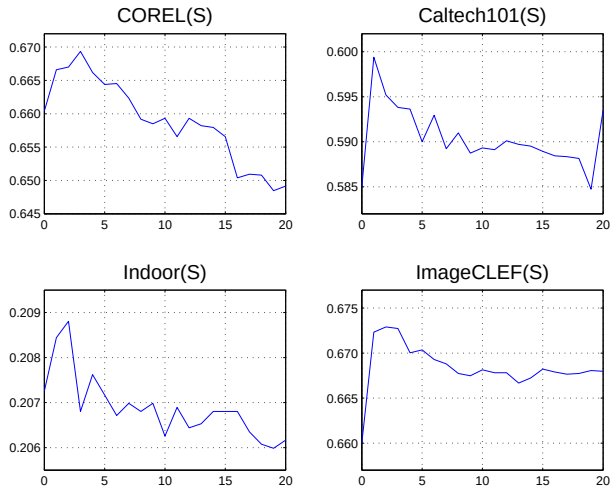


Fig. 4. Evaluation of the mAP (y-axis) of OMDL-LR w.r.t. the number of Nearest Neighbors (x-axis).

the mAP performance and the training time cost of OMDL-LR on four datasets, "COREL(S)", "Caltech101(S)", "Indoor(S)" and "ImageCLEF(S)", under the same experimental setting as the previous sections. The parameters for the OMDL-LR algorithm are set as follows: (i) d_{LR}, the dimensionality of the low-rank for all the models is set to 50, the same as the rank setting of r for the LOMDML algorithm; (ii) other hyper-parameters, including C_1, C_2, η and the number of nearest neighbors ("NN") for graph Laplacian, are determined by grid search on a separated validation set. Fig. 4 shows the mAP with respect to "NN" on each dataset.

From the comparison results in TABLE 6, we observed that LOMDML is even better than OMDL-LR in terms of the mAP performance. This may seem counterintuitive as OMDL-LR is a kernel-based approach. However, we conjecture that this is primarily because OMDL-LR fairly depends on a good selection of the underlying kernels and the parameters of the kernel functions. With carefully selected kernels, OMDL-LR would likely achieve better results. However, how to tune and find the best kernels is beyond the scope of this paper. In terms of training time cost, we observed that LOMDML is considerably more efficient than OMDL-LR. Similar to OMDML, the most computationally intensive step in OMDL-LR is the PSD projection, which can be $\mathcal{O}(r^3)$ for a dense matrix, thus the overall time complexity is $\mathcal{O}(T \times m \times r^3)$. In the above experiment, the dimensions of raw features range from 37 to 512, which are much smaller than $r^2 = 2500$. Thus, LOMDML consumes much less time than OMDL-LR.

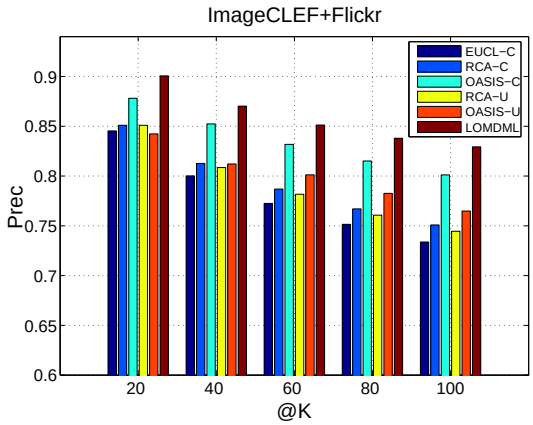


Fig. 5. Precision at Top-K on "ImageCLEF+Flickr"

5.9 Evaluation on the Large-scale Dataset

To examine its scalability, we apply the proposed algorithm on a large-scale image retrieval application on "ImageCLEF+Flickr", which has over one million images and 300K triplet training data. TABLE 7 shows the mAP performance of the five algorithms.

TABLE 7
Evaluation of mAP on the "ImageCLEF+Flickr" dataset.

Eucl-C	RCA-C	OASIS-C	RCA-U	OASIS-U	LOMDML
0.5766	0.6163	0.7161	0.6219	0.7028	0.7413

Clearly, our proposed algorithm OLMDML achieves the best mAP. Figure 5 presents the top-K precisions on ImageCLEF+Flickr. We can have the similar observation that our proposed methods significantly outperform the state of the art, in terms of precision. In short, the proposed algorithm significantly outperforms the state of the art, in terms of both mAP and retrieval accuracy performance measures.

5.10 Qualitative Comparison

Finally, to examine the qualitative retrieval performance, we randomly sample some query images from the query set, and compare the qualitative image similarity search by different algorithms. Figure 6 shows the comparison of retrieval results on "COREL" and "Caltech101" datasets using different algorithms. From the visual results, we can see that LOMDML generally returns more related results than the other baselines.

6 Conclusions

This paper investigated a novel family of online multi-modal distance metric learning (OMDML) algorithms for CBIR tasks by exploiting multiple types of features. We pinpointed some major limitations of traditional DML approaches in practice, and presented the online multi-modal DML method which simultaneously learns both the optimal distance metric on each individual feature space and the optimal combination of multiple metrics on different types of features. Further, we proposed the low-rank online multi-modal DML algorithm (LOMDML), which not only runs more efficiently and scalably, but also achieves the state-of-the-art performance among the competing algorithms in our experiments. Future work can

extend our framework in resolving other types of multimodal data analytics tasks beyond image retrieval.

ACKNOWLEDGEMENTS

This work was supported by Singapore MOE tier-1 research grant from Singapore Management University, Singapore.

REFERENCES

[1] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *Multimedia Computing, Communications and Applications, ACM Transactions on*, vol. 2, no. 1, pp. 1–19, 2006.

[2] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 11, pp. 1877–1890, 2008.

[3] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 8, pp. 1371–1384, 2008.

[4] A. K. Jain and A. Vailaya, "Shape-based retrieval: a case study with trademark image database," *Pattern Recognition*, no. 9, pp. 1369–1390, 1998.

[5] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth movers distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, p. 2000, 2000.

[6] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, 2000.

[7] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, New York, US, Jun. 17–22 2006, dCA.

[8] L. Si, R. Jin, S. C. Hoi, and M. R. Lyu, "Collaborative image retrieval via regularized metric learning," *ACM Multimedia Systems Journal*, vol. 12, no. 1, pp. 34–44, 2006.

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

[10] G. H. J. Goldberger, S. Roweis and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems*, 2005.

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

[12] A. Globerson and S. Roweis, "Metric learning by collapsing classes," in *Advances in Neural Information Processing Systems*, 2005.

[13] L. Yang, R. Jin, R. Sukthankar, and Y. Liu, "An efficient algorithm for local distance metric learning," in *Association for the Advancement of Artificial Intelligence*, 2006.

[14] A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognition*, vol. 29, pp. 1233–1244, 1996.

[15] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 8, pp. 837–842, 1996.

[16] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[17] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *ACM International Conference on Multimedia Information Retrieval*, 2007, pp. 197–206.

[18] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, 1999, pp. 1150–1157.

[19] R. S. Mohammad Norouzi, David Fleet, "Hamming distance metric learning," in *Advances in Neural Information Processing Systems*, 2012.

[20] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.

[21] H. Chang and D.-Y. Yeung, "Kernel-based distance metric learning for content-based image retrieval," *Image and Vision Computing*, vol. 25, no. 5, pp. 695–703, 2007.

[22] R. Salakhutdinov and G. Hinton, "Semantic hashing," *International Journal of Approximate Reasoning*, vol. 50, no. 7, pp. 969–978, Jul. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.ijar.2008.11.006

[23] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2011.235

[24] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *BMVC*, 2011, pp. 1–12.

[25] A. Joly and O. Buisson, "Random maximum margin hashing," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, Washington, DC, USA, 2011, pp. 873–880.

[26] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao, "Multiview metric learning with global consistency and local smoothness," *ACM Trans. on Intelligent Systems and Technology*, vol. 3, no. 3, p. 53, 2012.

[27] W. Di and M. Crawford, "View generation for multiview maximum disagreement based active learning for hyperspectral image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 50, no. 5, pp. 1942–1954, 2012.

[28] S. Akaho, "A kernel method for canonical correlation analysis," in *In Proceedings of the International Meeting of the Psychometric Society*. Springer-Verlag, 2001.

[29] J. D. R. Farquhar, H. Meng, S. Szedmak, D. R. Hardoon, and J. Shawe-taylor, "Two view learning: Svm-2k, theory and practice," in *Advances in Neural Information Processing Systems*. MIT Press, 2006.

[30] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *Proceedings of International Conference on Machine Learning*, 2003, pp. 11–18.

[31] J.-E. Lee, R. Jin, and A. K. Jain, "Rank-based distance metric learning: An application to image retrieval," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008.

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

[33] C. Domeniconi, J. Peng, and D. Gunopulos, "Locally adaptive metric nearest-neighbor classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1281 – 1285, 2002.

[34] P. Wu, S. C. H. Hoi, P. Zhao, and Y. He, "Mining social images with distance metric learning for automated image tagging," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 197–206.

[35] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 153–162.

[36] X. Gao, S. C. Hoi, Y. Zhang, J. Wan, and J. Li, "Soml: Sparse online metric learning with application to image retrieval," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[37] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman, "Online metric learning and fast similarity search," in *Advances in Neural Information Processing Systems*, 2008, pp. 761–768.

[38] R. Jin, S. Wang, and Y. Zhou, "Regularized distance metric learning: Theory and algorithm," in *Advances in Neural Information Processing Systems*, 2009, pp. 862–870.

[39] D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. He, and C. Miao, "Learning to name faces: a multimodal learning scheme for search-based face annotation," in *SIGIR*, 2013, pp. 443–452.

[40] H. Xia, P. Wu, and S. C. H. Hoi, "Online multi-modal distance learning for scalable multimedia retrieval," in *WSDM*, 2013, pp. 455–464.

[41] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[42] S. C. Hoi, J. Wang, and P. Zhao, "Libol: A library for online learning algorithms," *Journal of Machine Learning Research*, vol. 15, pp. 495–499, 2014. [Online]. Available: https://github.com/LIBOL

[43] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 7, pp. 551–585, 1958.

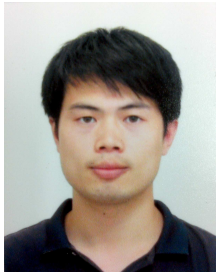
[44] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 551–585, 2006.

[45] M. Dredze, K. Crammer, and F. Pereira, "Confidence-weighted linear classification," in *Proceedings of International Conference on Machine Learning*, 2008, pp. 264–271.

[46] K. Crammer, A. Kulesza, and M. Dredze, "Adaptive regularization of weight vectors," in *Advances in Neural Information Processing Systems*, 2009, pp. 414–422.
[47] P. Zhao, S. C. H. Hoi, and R. Jin, "Double updating online learning," *Journal of Machine Learning Research*, vol. 12, pp. 1587–1615, 2011.
[48] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[49] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of International Conference on Machine Learning*, 2003, pp. 928–936.
[50] S. C. H. Hoi, J. Wang, P. Zhao, R. Jin, and P. Wu, "Fast bounded online gradient descent algorithms for scalable kernel-based online learning," in *ICML*, 2012.
[51] S. C. Hoi, R. Jin, P. Zhao, and T. Yang, "Online multiple kernel classification," *Machine Learning*, vol. 90, no. 2, pp. 289–316, 2013.
[52] Y. Freund and R. E. Schapire, "Adaptive game playing using multiplicative weights," *Games and Economic Behavior*, vol. 29, no. 1, pp. 79–103, 1999.
[53] Y. Li and P. M. Long, "The relaxed online maximum margin algorithm," in *Advances in Neural Information Processing Systems*, 1999, pp. 498–504.
[54] L. Bottou and Y. LeCun, "Large scale online learning," in *Advances in Neural Information Processing Systems*, 2003.
[55] S. C. Hoi, M. R. Lyu, and R. Jin, "A unified log-based relevance feedback scheme for image retrieval," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 4, pp. 509–204, 2006.
[56] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007.
[57] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[58] L. Yang, R. Jin, L. B. Mummert, R. Sukthankar, A. Goode, B. Zheng, S. C. H. Hoi, and M. Satyanarayanan, "A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 1, pp. 30–44, 2010.
[59] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 399–402.
[60] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Information fusion in multimedia information retrieval," *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics*, pp. 147–159, 2008.

Pengcheng Wu received his PhD degree from the School of Computer Engineering at the Nanyang Technological University, Singapore, and his bachelor degree from Xiamen University, P.R. China. He is currently a research fellow in the School of Information Systems, Singapore Management University. His research interests include multimedia information retrieval, machine learning and data mining.

Steven C. H. Hoi is currently an Associate Professor of the School of Information Sytems, Singapore Management Unviersity, Singapore. Prior to joining SMU, he was Associate Professor with Nanyang Technological University, Singapore. He received his Bachelor degree from Tsinghua University, P.R. China, in 2002, and his Ph.D degree in computer science and engineering from The Chinese University of Hong Kong, in 2006. His research interests are machine learning and data mining and their applications to multimedia information retrieval (image and video retrieval), social media and web mining, and computational finance, etc, and he has published over 150 refereed papers in top conferences and journals in these related areas. He has served as Associate Editor-in-Chief for Neurocomputing Journal, general co-chair for ACM SIGMM Workshops on Social Media (WSM'09, WSM'10, WSM'11), program co-chair for the fourth Asian Conference on Machine Learning (ACML'12), book editor for "Social Media Modeling and Computing", guest editor for ACM Transactions on Intelligent Systems and Technology (ACM TIST), technical PC member for many international conferences, and external reviewer for many top journals and worldwide funding agencies, including NSF in US and RGC in Hong Kong.



Peilin Zhao received his PhD from the School of Computer Engineering at the Nanyang Technological University, Singapore, in 2012 and his bachelor degree from Zhejiang University, Hangzhou, P.R. China, in 2008. His research interests are statistical machine learning, and data mining.

Chunyan Miao is an Associate Professor in the School of Computer Engineering at Nanyang Technological University (NTU). Her research focus is on infusing intelligent agents into interactive new media (virtual, mixed, mobile and pervasive media) to create novel experiences and dimensions in game design, interactive narrative and other real world agent systems. She has done significant research work her research areas and published many top quality international conference and journal papers.

Zhi-Yong Liu received his Bachelor degree of Engineering from Tianjin University in 1997, Master degree of Engineering from Chinese Academy of Sciences in 2000, and Ph.D degree from the Chinese University of Hong Kong in 2003. He is currently a professor at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, China. His research interests include image analysis, pattern recognition, machine learning and computer vision.

Fig. 6. Qualitative evaluation of top-5 retrieved images by different algorithms. For each block, the first image is the query, and the results from the first line to the sixth line represents “Eucl-C”, “RCA-C”, “OASIS-C”, “RCA-U”, “OASIS-U” and “LOMDML” respectively. The left column is from the “Corel” dataset and the right is from the “Caltech101” dataset.