SOML: Sparse Online Metric Learning with Application to Image Retrieval

Xingyu Gao[1,2,3], **Steven C.H. Hoi**[2], **Yongdong Zhang**[1], **Ji Wan**[1,2,3], **Jintao Li**[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, China
[2]School of Computer Engineering, Nanyang Technological University, Singapore
[3]University of Chinese Academy of Sciences, Beijing, China
chhoi@ntu.edu.sg, {gaoxingyu, zhyd, wanji, jtli}@ict.ac.cn

Abstract

Image similarity search plays a key role in many multimedia applications, where multimedia data (such as images and videos) are usually represented in high-dimensional feature space. In this paper, we propose a novel Sparse Online Metric Learning (SOML) scheme for learning sparse distance functions from large-scale high-dimensional data and explore its application to image retrieval. In contrast to many existing distance metric learning algorithms that are often designed for low-dimensional data, the proposed algorithms are able to learn sparse distance metrics from high-dimensional data in an efficient and scalable manner. Our experimental results show that the proposed method achieves better or at least comparable accuracy performance than the state-of-the-art non-sparse distance metric learning approaches, but enjoys a significant advantage in computational efficiency and sparsity, making it more practical for real-world applications.

Introduction

With the popularity of social media applications, recent years have witnessed an explosive growth of multimedia data on the Internet. For many real-world multimedia applications, image similarity search is a fundamental research task that has been actively studied for many years in several communities. The key challenges of this research are mainly twofold. The first is to design effective feature representation, and the second is to study effective and efficient distance/similarity functions over the feature space. For the feature representation, researchers in multimedia and computer vision have proposed a variety of effective features for content-based image retrieval (Smeulders et al. 2000; Rahmani et al. 2008) in the past decade. Examples include global features: color, texture and shape (Gevers and Smeulders 2000), and local features: SIFT feature descriptors (Lowe 1999; 2004; Mikolajczyk and Schmid 2005; Quelhas et al. 2007; Zhang et al. 2007) and SURF feature descriptors (Bay, Tuytelaars, and Gool 2006) as well as their Bag-of-Words (BoW) representation (Fergus et al. 2005; Wang, Zhang, and Li 2006; Bosch, Muñoz, and Marti 2007; Jiang et al. 2010; Jegou, Douze, and Schmid 2010; Wu, Hoi, and Yu 2010; Luo, Wei, and Lai 2011). For distance/similarity functions, a variety of schemes have been proposed in multimedia and computer vision. The most widely used examples include Cosine similarity and Euclidean distance, both of which assume a rigid similarity or distance function in some vector space which are often optimal for the applications.

Instead of using rigid distance/similarity functions, Distance Metric Learning (DML) techniques (Yang and Jin 2006) have been actively explored to optimize distance metrics in various applications, such as image retrieval (Yang et al. 2010; Hoi, Liu, and Chang 2008), image annotation (Wu et al. 2011), and pose and facial expression matching (Zhai et al. 2012), etc. Specifically, the goal of DML typically is to optimize the Mahalanobis distance or parametric bi-linear similarity function between two data instances, which are mathematically expressed as respectively:

$$\mathcal{D}_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)} \quad (1)$$

$$\mathcal{S}_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{M} \mathbf{x}_j \quad (2)$$

where $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^m$ and $\mathbf{M} \in \mathbb{R}^{m \times m}$ must be positive semi-definite to satisfy the properties of a distance metric. A variety of DML learning algorithms have been proposed to find an optimal Mahalanobis matrix $\mathbf{M}$ from the training data. However, most existing DML algorithms are designed to learn distance metrics in low-dimensional feature space, and are computationally inefficient and non-scalable for high-dimensional data, making them unsuitable for real-world image retrieval applications.

To tackle the above challenge, we propose a novel Sparse Online Metric Learning (SOML) scheme for learning sparse distance metrics from high-dimensional data and explore its application to image retrieval. In particular, we propose to tackle the problem of online distance metric learning in high-dimensional space by employing recent advances in Sparse Online Learning techniques in machine learning (Langford, Li, and Zhang 2009; Xiao 2010), which are able to learn a sparse distance metric from pairwise high-dimensional training data in a highly efficient and scalable manner. We apply the proposed SOML technique to optimize the (high-dimensional) Bag-of-Words (BoW) representation of images in content-based image retrieval, and conduct extensive experiments by comparing with some

state-of-the-art solutions, in which the proposed SOML method is able to achieve comparable or sometimes even better accuracy, but enjoys a significant advantage in sparsity and computational efficiency, making it more practical for real-world applications. In summary, the main contributions of this paper include:

- We present a novel Sparse Online Metric Learning (SOML) method for learning sparse distance metrics from large high-dimensional data.
- We propose two sparse online metric learning algorithms and explore their applications for optimizing the high-dimensional BoW representation in image retrieval.
- We conduct extensive experiments by comparing the proposed algorithms with the state-of-the-art method for optimizing the BoW representation in image retrieval.

The rest of this paper is organized as follows. We first briefly review related work, and then present the problem formulation and the proposed algorithms for Sparse Online Metric Learning (SOML) with application to image retrieval. We further discuss our experimental results, and finally conclude this paper.

Related Work

Our work lies in the intersection of machine learning and multimedia information retrieval. In this section, we briefly review two major categories of related work in machine learning, multimedia and computer vision.

Distance Metric Learning

Our work is closely related to Distance Metric Learning (DML), which has been extensively studied in machine learning, multimedia and computer vision communities (Yang and Jin 2006). A variety of DML algorithms have been proposed by following different settings and methodologies across different communities. In terms of training data formats, most existing DML work can be generally grouped into two major categories: (i) learning distance metrics directly from explicit class labels (Weinberger and Saul 2009) which are common for generic data classification tasks, and (ii) learning distance metrics from side information (either pairwise (Hoi, Liu, and Chang 2008; Hoi et al. 2006) or triplet constraints (Chechik et al. 2010; Wu et al. 2013)), which are common for multimedia retrieval applications (Hoi, Lyu, and Jin 2006).

In terms of learning methodology, most existing DML methods often adopt batch machine learning approaches. The major limitation of such learning methodology is that the metric has to be re-trained from scratch whenever there is new training data. In recent years, some emerging DML studies have explored online learning techniques to tackle DML tasks (Jain et al. 2008; Chechik et al. 2010). Our work also follows the online learning methodology (Hoi, Wang, and Zhao 2014) to tackle DML tasks.

Despite a variety of DML techniques proposed in the literature, one common issue with the existing DML approaches is that they often learn a full matrix from low-dimensional data or sometimes learn a dense diagonal matrix from high-dimensional data. Such approaches often result in complicated optimization tasks, making them hardly scalable for very high-dimensional data. Besides, learning a full matrix or a dense diagonal matrix for distance metrics also will lead to high computational cost when calculating the distance in the final applications. Unlike the existing DML approaches, our goal is to study a highly efficient and scalable online learning scheme for learning sparse distance metrics from very high-dimensional data.

Sparse Online Learning

Our work is also related to sparse online learning in machine learning (Langford, Li, and Zhang 2009; Duchi and Singer 2009), which aims to induce sparsity in the model learned by an online learner. Mathematically, sparse online learning can be formulated as formal online optimization tasks with convex objective functions and some sparsity-promoting regularizer (Duchi and Singer 2009). A variety of algorithms have been proposed to resolve such online optimization tasks efficiently. In terms of different optimization principles, there are two major groups of sparse online learning algorithms.

The first group follows the general idea of subgradient descent with truncation also known as the Truncated Gradient (TG) for short. For example, FOBOS (Duchi and Singer 2009) adopts a traditional subgradient descent step followed by an instantaneous minimization that keeps close to the update with a sparsity-promoting penalty. By arguing that the truncation at every iteration is too aggressive, an improved TG method (Langford, Li, and Zhang 2009) is proposed, which truncates coefficients every step only when the coefficients exceed a predefined threshold. The second group of algorithms is based on the idea of Dual Averaging (DA) methods for sparsity-inducing online optimization (Xiao 2010). For instance, (Xiao 2010) extends the simple dual averaging scheme by proposing the regularized dual averaging (RDA) algorithm, which uses a much more aggressive truncation threshold and is able to generate significantly sparser solutions.

Despite active study, most existing work focuses on learning classifiers for online classification tasks. In this work, we extend sparse online learning techniques for solving DML tasks, where the training data is given in the form of triplet constraints. In our framework, we explore both truncated gradient and dual averaging based algorithms to tackle our sparse online metric learning problem.

Sparse Online Metric Learning

Problem Formulation

We address the fundamental problem of distance metric learning from side information (pairwise or triplet image relationship) towards image retrieval applications. To formulate the metric learning task, we denote the similarity function between any two images $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^m$ by notation $\mathcal{S}(\mathbf{x}_i, \mathbf{x}_j)$ and assume a collection of training data instances are given sequentially in the form of triplet instances $\{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-), i = 1, \cdots, n\}$, where each triplet instance

indicates the triplet relationship of three images, i.e., image $\mathbf{x}_i$ is more similar to image $\mathbf{x}_i^+$ than image $\mathbf{x}_i^-$, and n is the total number of triplets. The goal is to learn a similarity function $\mathcal{S}(\cdot,\cdot)$ which can produce the similarity values always satisfying the triplet constraints as follows:

$$\mathcal{S}(\mathbf{x}_i, \mathbf{x}_i^+) \geq 1 + \mathcal{S}(\mathbf{x}_i, \mathbf{x}_i^-), \quad \forall \mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^- \in \mathcal{X} \tag{3}$$

where 1 is a margin constant to ensure that $\mathcal{S}(\mathbf{x}_i, \mathbf{x}_i^+)$ is sufficiently larger than $\mathcal{S}(\mathbf{x}_i, \mathbf{x}_i^-)$.

In this paper, we aim to explore DML techniques for image retrieval applications, where images are often represented as a Bag-of-Words (BoW) feature vector in high-dimensional space. Thus, we consider the problem of online metric learning with a linear similarity function $\mathcal{S}$ defined as:

$$\mathcal{S}_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbf{x}_i^T \mathbf{M} \mathbf{x}_j \tag{4}$$

where $\mathbf{M} \in \mathbb{R}^{m \times m}$. It is not difficult to see that the above similarity function reduces to cosine similarity when choosing $\mathbf{M}$ as an identity matrix and assuming instances are of unit norm.

Given the above similarity function and the constraint in (3), we can formulate the problem of distance metric learning as a convex optimization task

$$\min_{\mathbf{M}} \sum_{i=1}^{n} \mathcal{L}((\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-); \mathbf{M}) + \lambda r(\mathbf{M}) \tag{5}$$

where $r(\mathbf{M})$ is some convex regularization term (e.g., sparsity-promoting regularizer) that limits the model complexity, $\lambda > 0$, and the loss function $\mathcal{L}$ is based on the hinge loss defined as:

$$\mathcal{L}((\mathbf{x}_i,\mathbf{x}_i^+,\mathbf{x}_i^-);\mathbf{M}) = \max(0, 1 - \mathcal{S}_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_i^+) + \mathcal{S}_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_i^-)) \tag{6}$$

Minimizing the above loss is equivalent to minimizing the violations on the constraints defined in (3).

The above optimization is a batch learning formulation with a full matrix $\mathbf{M}$ of space complexity $O(m^2)$, which poses a huge challenge when handling high-dimensional data. In order to deal with very high-dimensional image data, we simplify the DML problem by considering the metric defined by a diagonal matrix, i.e., $\mathbf{M} = \mathrm{diag}(\mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^m$. We can rewrite the loss function $\mathcal{L}$ into:

$$\mathcal{L}((\mathbf{x}_i,\mathbf{x}_i^+,\mathbf{x}_i^-);\mathbf{w}) = \max(0, 1 - \mathcal{S}_{\mathbf{w}}(\mathbf{x}_i,\mathbf{x}_i^+) + \mathcal{S}_{\mathbf{w}}(\mathbf{x}_i,\mathbf{x}_i^-)) \tag{7}$$

where $\mathcal{S}_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbf{x}_i^T \mathrm{diag}(\mathbf{w}) \mathbf{x}_j$. We can now give the online learning formulation as follows.

By following online learning settings (Hoi, Wang, and Zhao 2014), we assume a triplet instance $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$ is received at every step $t = 1, \cdots, n$. The goal of Sparse Online Metric Learning (SOML) is to sequentially update the metric $M = \mathrm{diag}(\mathbf{w})$ by solving the following online optimization:

$$\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}) + \lambda r(\mathbf{w}) \tag{8}$$

where $r(\mathbf{w})$ is a sparsity-promoting regularizer which is based on the ℓ_1-regularizer, i.e., $r(\mathbf{w}) = \|\mathbf{w}\|_1$ in our approach. In the following, we present two efficient algorithms to tackle the above optimization task of online sparse metric learning for handling very high-dimensional data.

SOML-TG: Sparse Online Metric Learning via Truncated Gradient Algorithm

We first attempt to solve the SOML problem by exploring the Truncated Gradient (TG) based technique (Langford, Li, and Zhang 2009), which extends the online gradient descent with truncation tricks.

Specifically, consider an online optimization with the objective function in (7), a simple online gradient descent (OGD) method for ℓ_1-regularization makes an update:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \nabla \mathcal{L}_{\mathbf{w}_t} - \eta \lambda \mathrm{sgn}(\mathbf{w}_t) \tag{9}$$

where $\nabla \mathcal{L}_{\mathbf{w}_t}$ is a sub-gradient of $\mathcal{L}$ with respect to $\mathbf{w}_t$. $\eta > 0$ is a learning rate parameter, and $\lambda > 0$ is a regularization parameter. This method however does not generate sparse weights online.

In order to produce sparse weights at every online step, we extend the OGD rule by applying the Truncated Gradient approach which performs the following truncation update:

$$\mathbf{w}_{t+1} \quad \leftarrow \quad T_1(\mathbf{w}_t - \eta \nabla \mathcal{L}_{\mathbf{w}_t}, \eta \lambda_t) \tag{10}$$

where $\lambda \geq 0$ and η is the learning rate, and $T_1(\mathbf{v}, \alpha) = [T_1(v_1, \alpha), T_1(v_2, \alpha), \cdots, T_1(v_m, \alpha)]$ is a truncation function in which each dimension is defined as

$$T_1(v_j, \alpha) = \begin{cases} \max(0, v_j - \alpha), & \text{if } v_j \geq 0 \\ \min(0, v_j + \alpha), & \text{otherwise} \end{cases} \tag{11}$$

By taking the specific form of $\nabla \mathcal{L}_{\mathbf{w}_t}$, we have

$$\mathbf{w}_{t+1} \leftarrow T_1(\mathbf{w}_t - \eta[\mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-)], \eta \lambda_t) \tag{12}$$

where $\odot$ denotes an elementwise product of two vectors. The above update tries to promote sparsity for the OGD solution $\mathbf{w}_t - \eta \nabla \mathcal{L}_{\mathbf{w}_t}$ by performing truncation with threshold $\eta \lambda_t$. Finally, Algorithm 1 summarizes the details of the proposed Sparse Online Metric Learning via Truncated Gradient (SOML-TG) algorithm.

Algorithm 1 SOML-TG—Sparse Online Metric Learning via Truncated Gradient

Input: Training Triplets: $(\mathbf{x}, \mathbf{x}_t^+, \mathbf{x}_t^-),\ t = 1, \ldots, n$.
Output: The weight vector $\mathbf{w}$.

Initialize $\mathbf{w}_1 = 0$; $\alpha = \eta\lambda$
repeat
  Receive a triplet instance $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$,
  Suffer loss $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t)$ measured by (7)
  if $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t) > 0$ **then**
    $\mathbf{v} = \mathbf{w}_t - \eta[\mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-)]$;
    for j=1 to m **do**
      if $\mathbf{v}_j \geq 0$ **then**
        $\mathbf{w}_{t+1,j} = \max(0, \mathbf{v}_j - \alpha)$;
      else
        $\mathbf{w}_{t+1,j} = \min(0, \mathbf{v}_j + \alpha)$;
      end if
    end for
  end if
until CONVERGENCE

SOML-DA: Sparse Online Metric Learning via Dual Averaging Algorithm.

Our second solution is to explore Nesterov's Dual Averaging (DA) method (Nesterov 2009) and its extensions (Xiao 2010) to tackle the sparse online metric learning problem, which attempts to exploit all past subgradients of the loss function and the whole regularization term, instead of only its subgradient by the truncated gradient approaches.

Specifically, when receiving a triplet instance $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$ at each online step, we update the weight vector by exploring a regularized dual averaging method with ℓ_1-regularization as follows:

$$\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \frac{1}{t}\sum_{i=1}^{t}\langle \nabla\mathcal{L}_{\mathbf{w}_i}, \mathbf{w}\rangle + \lambda_t\|\mathbf{w}\|_1 + \frac{\gamma}{2\sqrt{t}}\|\mathbf{w}\|^2 \quad (13)$$

where $\nabla\mathcal{L}_{\mathbf{w}_i}$ is a subgradient of $\mathcal{L}$ at the i-th online step, and $\frac{1}{2}\|\mathbf{w}\|^2$ is an auxiliary strongly convex function. λ_t is a truncating threshold $\lambda_t = \lambda + \frac{\gamma\rho}{\sqrt{t}}$, and $\lambda \geq 0$, $\gamma > 0$ and $\rho \geq 0$ are sparsity-promoting parameters. $\frac{\gamma}{\sqrt{t}}$ is a nonnegative and decreasing input sequence to ensure that the impact by the auxiliary function decreases with time. In online implementations, we maintain an average gradient $\bar{\nabla}_t$ at the t-th step:

$$\bar{\nabla}_t = \frac{t-1}{t}\bar{\nabla}_{t-1} + \frac{1}{t}\nabla_t\mathcal{L}_{\mathbf{w}_t} \quad (14)$$

$$= \frac{t-1}{t}\bar{\nabla}_{t-1} + \frac{1}{t}\mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-) \quad (15)$$

Using the above notation, we can derive the closed-form solution of $\mathbf{w}_{t+1} = [\mathbf{w}_{t+1}^{(1)}, \ldots, \mathbf{w}_{t+1}^{(m)}]$ for optimizing (13) as:

$$\mathbf{w}_{t+1}^{(i)} = \begin{cases} 0, & \text{if } |\bar{\nabla}_t{}^{(i)}| \leq \lambda_t \\ -\frac{\sqrt{t}}{\gamma}\big(\bar{\nabla}_t{}^{(i)} - \lambda_t \mathrm{sgn}(\bar{\nabla}_t{}^{(i)})\big), & \text{otherwise} \end{cases} \quad (16)$$

where λ_t is a truncating threshold $\lambda_t = \lambda + \frac{\gamma\rho}{\sqrt{t}}$, and $\rho \geq 0$ is the sparsity-promoting parameter. Finally, Algorithm 2 summarizes the details of the proposed Sparse Online Metric Learning via Dual Averaging (SOML-DA) algorithm.

Algorithm 2 SOML-DA—Sparse Online Metric Learning via Dual Averaging Algorithm

Input:
Training Triplets: $(\mathbf{x}, \mathbf{x}_t^+, \mathbf{x}_t^-),\ t = 1, \ldots, n$
Input parameters: $\gamma > 0$, $\rho \geq 0$
Output: The final weight vector $\mathbf{w}_{n+1}$.
Initialize $\mathbf{w}_1 = \mathbf{0}$, $\bar{\nabla}_0 = \mathbf{0}$
repeat
  Receive a triplet instance $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$,
  Suffer loss $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t)$ measured by (7)
  Compute $\bar{\nabla}_t = \frac{t-1}{t}\bar{\nabla}_{t-1} + \frac{1}{t}\mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-)$
  Compute $\lambda_t = \lambda + \gamma\rho/\sqrt{t}$
  if $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t) > 0$ **then**
    for j=1 to m **do**
      if $|\bar{\nabla}_t{}^{(j)}| \leq \lambda_t$ **then**
        $\mathbf{w}_{t+1}^{(j)} = 0$;
      else
        $\mathbf{w}_{t+1}^{(j)} = -\frac{\sqrt{t}}{\gamma}\big(\bar{\nabla}_t{}^{(j)} - \lambda_t \mathrm{sgn}(\bar{\nabla}_t{}^{(j)})\big)$;
      end if
    end for
  end if
until CONVERGENCE

Experiments

In our experiments, we investigate the application of the proposed sparse online metric learning technique for improving the Bag-of-Words (BoW) representation in image retrieval tasks. In the following, we first introduce the experimental testbed and setup, followed by discussing the detailed experimental results.

Experimental Testbed and Setup

In order to examine the efficacy of the proposed sparse online metric learning scheme, we compare the following schemes for image retrieval in our experiments:

- TF-IDF: the commonly used TF-IDF scheme for weighing the BoW representation (Baeza-Yates, Ribeiro-Neto, and others 1999);
- QPAO: a state-of-the-art codebook learning approach (Cai, Yan, and Mikolajczyk 2010) which formulates it as quadratic programming (QP) and adopts Alternating Optimization (AO) to solve it.
- OASIS: Online Algorithm for Scalable Image Similarity (OASIS) (Chechik et al. 2010), a state-of-the-art algorithm for online metric learning.
- SOML: the two proposed sparse online metric learning algorithms: SOML-TG and SOML-DA, which denote by "S-TG" and "S-DA" for short.

Following previous studies, we adopt the "Oxford5K" image dataset, a well-known public dataset for image retrieval benchmarks. This dataset contains a total of 5,062 images for 11 Oxford landmarks with manually annotated ground truth. We follow the same experimental settings used in previous studies, where 5 images per landmark are used for each query. The mean Average Precision (mAP) is employed as the performance metric for evaluating the retrieval results. We learn distance/similarity metrics for each landmark with 7 randomly selected positive images and 500 negative images, which generates a total of 21,000 ($= 7 \times 6 \times 500$) triplet instances. The remaining 4,555 images are used for testing/retrieval. We evaluate the list of compared algorithms on 7 landmarks out of 11 and exclude another 4 landmarks because they simply have too few training examples to learn by the algorithms. The same setting was also adopted by the previous codebook learning study in (Cai, Yan, and Mikolajczyk 2010). Finally, for the BoW representation of images, we use SIFT for feature descriptors, and Approximate K-means (AKM) clustering to generate different-sized codebooks in three scales: 10,000, 100,000, and 1-million.

For parameter settings, we set the parameters for the three different-sized codebooks (10,000, 100,000, and 1-million) are: $L = 0$, $M = 10^4$, 50 subsets; $L = 0$, $M = 10^5$, 50 subsets; and $L = 0$, $M = 10^6$, 500 subsets for QPAO (Cai, Yan, and Mikolajczyk 2010). For OASIS, the parameters are $C = 0.1$ and 10^5 training steps with different-sized codebooks. For SOML-TG algorithm, we set parameters $\eta = 1$, $\lambda = 10^{-5}$ for all the codebooks respectively. For SOML-DA algorithm, we set parameters $\gamma = 10^{-4}$, $\rho = 1$, and $\lambda = 10^{-6}$ with the three different-sized codebooks respectively.

Experimental Results

We first evaluate the retrieval quality of different schemes, and then evaluate the sparsity of the learned weights as well as the computational efficiency of the different algorithms.

Evaluation of Mean Average Precision Table 1, Table 2, and Table 3 show the evaluation of mAP performance by different schemes. We draw several observations from the experimental results.

First of all, we observe that all learning-based schemes are able to outperform the TF-IDF scheme without learning for most cases. This shows the efficacy and importance of optimizing the BoW representation by applying machine learning techniques in exploiting side info/training data.

Table 1: Comparison of mean Average Precision (%) on Oxford5K dataset with 10,000-sized codebook.

Category	TF-IDF	QPAO	OASIS	S-TG	S-DA
all souls	40.60	**57.42**	52.72	56.68	45.50
ashmolean	30.66	30.02	33.03	30.79	**35.90**
bodleian	30.11	68.28	65.13	64.66	**74.53**
christ church	46.35	45.79	43.41	**53.43**	52.42
hertford	31.16	**51.47**	42.18	44.30	35.93
magdalen	5.92	8.86	9.34	12.12	**17.55**
radcliffe camera	52.22	**82.44**	75.12	80.68	74.71
mAP	33.86	**49.18**	45.85	48.95	48.08

Table 2: Comparison of mean Average Precision (%) on Oxford5K dataset with 100,000-sized codebook.

Category	TF-IDF	QPAO	OASIS	S-TG	S-DA
all souls	58.17	**93.92**	75.22	91.58	88.70
ashmolean	44.96	42.78	**47.68**	40.15	41.12
bodleian	49.06	**86.02**	71.36	83.07	83.42
christ church	52.08	**70.74**	50.97	59.73	59.04
hertford	53.51	**63.93**	57.75	63.41	60.44
magdalen	11.29	10.99	12.96	9.63	**18.42**
radcliffe camera	70.51	**82.19**	76.92	76.69	76.43
mAP	48.51	**64.37**	56.13	60.61	61.08

Table 3: Comparison of mean Average Precision (%) on Oxford5K dataset with 1 million-sized codebook.

Category	TF-IDF	QPAO	OASIS	S-TG	S-DA
all souls	53.96	62.99	55.03	62.71	**63.57**
ashmolean	**53.86**	48.77	53.45	48.63	51.53
bodleian	66.88	**90.21**	70.57	84.17	83.47
christ church	56.67	**65.36**	57.34	61.10	61.10
hertford	72.00	68.66	75.36	79.24	**82.78**
magdalen	18.98	15.63	**19.24**	8.79	9.77
radcliffe camera	64.43	62.94	**65.04**	58.03	61.85
mAP	55.25	**59.23**	56.58	57.53	59.15

Second, we found that the batch learning approach, QPAO, achieved overall better retrieval performance than OASIS that is an online metric learning scheme. This is not too surprising since QPAO solves a batch optimization which thus might get better solution, while all online algorithms (OASIS and our algorithms) only learn from a single pass of the triplet instance sequence.

Further, by examining the two proposed algorithms (S-TG and S-DA), we found that their overall retrieval performance is better than OASIS, which indicates the proposed online metric learning algorithm with sparsity is potentially more effective than the existing online metric learning without exploiting sparsity. Moreover, by comparing with QPAO, we found that our algorithms are in general fairly comparable for most cases, and sometimes even better than QPAO (e.g., on the scenario with the 10,000-sized codebook). This encouraging result validates the efficacy of the proposed learning scheme for improving the BoW performance.

Finally, the two proposed algorithms (S-TG and S-DA) achieve very comparable retrieval performance in which S-DA tends to slightly outperform S-TG.

Evaluation of Sparsity of the Learned Weights The sparsity of BoW plays a critical role for large-scale content-based image retrieval systems, especially in the phases of image indexing and retrieval. A sparse BoW model not only can speed up the indexing and retrieval process, but also can save a significant amount of storage cost. Below we measure the sparsity of the learned weights by different algorithms, i.e., the number of zero values in the learned weight vectors.

Table 4: Comparison of sparsity (%) of learned weights by different approaches with 10,000-sized codebook.

Category	QPAO	OASIS	S-TG	S-DA
all souls	44.99	0.00	22.36	**64.54**
ashmolean	40.16	0.00	25.10	**74.49**
bodleian	35.80	0.00	77.43	**93.43**
christ church	31.91	0.00	32.08	**76.92**
hertford	43.19	0.00	22.87	**61.82**
magdalen	47.59	0.00	19.62	**52.61**
radcliffe camera	43.37	0.00	40.48	**75.53**

Table 5: Comparison of sparsity (%) of learned weights by different approaches with 100,000-sized codebook.

Category	QPAO	OASIS	S-TG	S-DA
all souls	0.02	0.00	90.64	**97.82**
ashmolean	0.02	0.00	86.37	**98.11**
bodleian	0.02	0.00	95.25	**98.82**
christ church	0.01	0.00	91.42	**98.87**
hertford	0.00	0.00	92.99	**97.56**
magdalen	0.04	0.00	82.26	**97.04**
radcliffe camera	0.01	0.00	93.47	**97.18**

Table 4, 5 and 6 show the sparsity evaluation of the learned weights by different learning approaches with 10,000-sized, 100,000-sized, and 1 million-sized codebook. Note that we use a small threshold 1E-10 to check if a value is zero or not. We can draw some observations from the results.

First, we found that OASIS fails to produce sparse weights for most cases, especially for large-sized code-

Table 6: Comparison of sparsity rate (%) of learned weights by different approaches with 1 million-sized codebook.

Category	QPAO	OASIS	S-TG	S-DA
all souls	0.00	0.00	99.51	**99.88**
ashmolean	0.00	0.00	99.30	**99.96**
bodleian	0.00	0.00	99.65	**99.85**
christ church	0.00	0.00	99.67	**99.97**
hertford	0.00	0.00	99.77	**99.90**
magdalen	0.00	0.00	99.06	**99.96**
radcliffe camera	0.00	0.00	99.51	**99.82**

books. QPAO is able to produce reasonably sparse weights on the 10,000-sized codebook, but also fails when the codebook size is large. By contrast, the two proposed algorithms (S-TG and S-DA) are able to produce sparse weights for all the cases. In particular, it seems the larger the codebook size, the higher the sparsity achieved by the proposed algorithms. Finally, by comparing the two proposed algorithms themselves, S-DA generally achieves better sparsity than S-TG for most cases primarily because it exploits all past subgradients and thus is a better approach for achieving sparsity.

Evaluation of Computational Cost Our last experiment is to evaluate the empirical computational cost of the different schemes. Table 7, Table 8 and Table 9 show the comparisons of training time costs by different learning schemes on three different-sized codebooks.

Table 7: Evaluation of training time cost (seconds) by different schemes with 10,000-sized codebook.

Category	QPAO	OASIS	S-TG	S-DA
all souls	2.98×10^3	17.91	8.83	**7.85**
ashmolean	2.51×10^3	17.34	6.61	**6.46**
bodleian	2.94×10^3	14.64	**8.51**	8.92
christ church	1.89×10^3	13.77	3.71	**3.56**
hertford	2.74×10^3	18.27	6.95	**5.58**
magdalen	2.49×10^3	17.75	6.23	**5.69**
radcliffe camera	2.93×10^3	16.82	**7.09**	7.69

Table 8: Evaluation of training time cost (seconds) by different schemes with 100,000-sized codebook.

Category	QPAO	OASIS	S-TG	S-DA
all souls	3.03×10^3	74.01	2.38	**1.76**
ashmolean	2.63×10^3	73.03	1.64	**1.14**
bodleian	3.89×10^3	68.48	3.57	**2.92**
christ church	1.95×10^3	69.08	0.99	**0.92**
hertford	2.87×10^3	73.78	2.10	**1.54**
magdalen	3.05×10^3	74.59	1.91	**1.26**
radcliffe camera	2.97×10^3	75.45	2.68	**1.81**

We can draw some observations from the experimental results. First of all, we can see that QPAO is the least efficient algorithm due to its QP formulation. Although QPAO has solved the QP problem by an efficient alternative optimization scheme, it remains inefficient when handling very high-dimensional data (e.g., 1-million scale). Second, OASIS is far more efficient than QPAO on relatively lower dimensional space since OASIS is an online algorithm whose time complexity is in general linear with respect to the sample size and dimensionality. However, when handling very high-dimensional data (e.g., on the 1-million sized codebook), OASIS becomes inefficient as the dimensionality plays a dominating factor.

Table 9: Evaluation of training time cost (seconds) by different schemes with 1 million-sized codebook.

Category	QPAO	OASIS	S-TG	S-DA
all souls	2.38×10^4	1.26×10^3	**1.34**	1.40
ashmolean	2.44×10^4	1.25×10^3	1.14	**1.01**
bodleian	1.94×10^4	1.24×10^3	2.24	**2.02**
christ church	2.17×10^4	1.25×10^3	0.99	**0.94**
hertford	2.19×10^4	1.26×10^3	**1.65**	1.72
magdalen	2.28×10^4	1.26×10^3	**1.05**	1.09
radcliffe camera	2.22×10^4	1.26×10^3	**1.65**	1.72

By contrast, the proposed algorithms are far more efficient and scalable than both QPAO and OASIS due to the proposed sparse online learning strategy. Finally, unlike the other two algorithms, it is interesting observe that increasing the dimensionality does not lead to increasing the time cost of the two proposed algorithms. This seems counter-intuitive but is not difficult to explain. This is because our algorithms always learn sparse weights in the online learning processes, and thus the time complexity of our algorithm depends on how many non-zero elements are in the training data instead of the dimensionality. This encouraging result again validates the efficiency and advantage of the proposed sparse online metric learning technique for large-scale real-world applications.

Conclusions

This paper presented a novel Sparse Online Metric Learning (SOML) scheme, aiming to make metric learning efficient and practical for handling very high-dimensional data. In particular, we explored the recent advances of sparse online learning for resolving the proposed distance metric learning problem, and presented two specific Sparse Online Metric learning (SOML) algorithms based on two different optimization solutions: truncated gradient based and dual averaging based algorithms. We investigate the application of the proposed SOML technique for improving the sparse and high-dimensional Bag-of-Words representation in image retrieval tasks. Our empirical results showed that the proposed SOML method is able to achieve comparable retrieval results in comparison to state-of-the-art approaches, but enjoys a significant gain in terms of model sparsity and computational cost, making our technique more practical for real-world applications. Future work will improve the current scheme by exploring more advanced sparse online learning techniques.

Acknowledgments

This work was supported in part by MOE Tier 1 Grant (RG33/11), National High Technology Research and Development Program of China (2014AA015202), National Key Technology Research and Development Program of China (2012BAH39B02), National Nature Science Foundation of China (61172153, 61100087).

References

Baeza-Yates, R.; Ribeiro-Neto, B.; et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Bay, H.; Tuytelaars, T.; and Gool, L. J. V. 2006. Surf: Speeded up robust features. In *ECCV (1)*, 404–417.

Bosch, A.; Muñoz, X.; and Marti, R. 2007. Which is the best way to organize/classify images by content? *Image Vision Comput.* 25(6):778–791.

Cai, H.; Yan, F.; and Mikolajczyk, K. 2010. Learning weights for codebook in image classification and retrieval. In *CVPR*, 2320–2327.

Chechik, G.; Sharma, V.; Shalit, U.; and Bengio, S. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research* 11:1109–1135.

Duchi, J., and Singer, Y. 2009. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research* 10:2899–2934.

Fergus, R.; Li, F.-F.; Perona, P.; and Zisserman, A. 2005. Learning object categories from google's image search. In *ICCV*, 1816–1823.

Gevers, T., and Smeulders, A. W. 2000. Pictoseek: Combining color and shape invariant features for image retrieval. *Image Processing, IEEE Transactions on* 9(1):102–119.

Hoi, S. C. H.; Liu, W.; Lyu, M. R.; and Ma, W.-Y. 2006. Learning distance metrics with contextual constraints for image retrieval. In *CVPR (2)*, 2072–2078.

Hoi, S. C. H.; Liu, W.; and Chang, S.-F. 2008. Semi-supervised distance metric learning for collaborative image retrieval. In *CVPR*.

Hoi, S. C. H.; Lyu, M. R.; and Jin, R. 2006. A unified log-based relevance feedback scheme for image retrieval. *IEEE Trans. Knowl. Data Eng.* 18(4):509–524.

Hoi, S. C.; Wang, J.; and Zhao, P. 2014. Libol: A library for online learning algorithms. *Journal of Machine Learning Research* 15:495–499.

Jain, P.; Kulis, B.; Dhillon, I. S.; and Grauman, K. 2008. Online metric learning and fast similarity search. In *NIPS*, 761–768.

Jegou, H.; Douze, M.; and Schmid, C. 2010. Improving bag-of-features for large scale image search. *International Journal of Computer Vision* 87(3):316–336.

Jiang, Y.-G.; Yang, J.; Ngo, C.-W.; and Hauptmann, A. G. 2010. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia* 12(1):42–53.

Langford, J.; Li, L.; and Zhang, T. 2009. Sparse online learning via truncated gradient. *Journal of Machine Learning Research* 10:777–801.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *ICCV*, 1150–1157.

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

Luo, H.-L.; Wei, H.; and Lai, L. L. 2011. Creating efficient visual codebook ensembles for object categorization. *IEEE Transactions on Systems, Man, and Cybernetics, Part A* 41(2):238–253.

Mikolajczyk, K., and Schmid, C. 2005. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(10):1615–1630.

Nesterov, Y. 2009. Primal-dual subgradient methods for convex problems. *Mathematical programming* 120(1):221–259.

Quelhas, P.; Monay, F.; Odobez, J.-M.; Gatica-Perez, D.; and Tuytelaars, T. 2007. A thousand words in a scene. *IEEE Trans. Pattern Anal. Mach. Intell.* 29(9):1575–1589.

Rahmani, R.; Goldman, S. A.; Zhang, H.; Cholleti, S. R.; and Fritts, J. E. 2008. Localized content-based image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 30(11):1902–1912.

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. Pattern Anal. Mach. Intell.* 22(12):1349–1380.

Wang, G.; Zhang, Y.; and Li, F.-F. 2006. Using dependent regions for object categorization in a generative framework. In *CVPR (2)*, 1597–1604.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10:207–244.

Wu, P.; Hoi, S. C. H.; Zhao, P.; and He, Y. 2011. Mining social images with distance metric learning for automated image tagging. In *WSDM*, 197–206.

Wu, P.; Hoi, S. C. H.; Xia, H.; Zhao, P.; Wang, D.; and Miao, C. 2013. Online multimodal deep similarity learning with application to image retrieval. In *ACM Multimedia*, 153–162.

Wu, L.; Hoi, S. C. H.; and Yu, N. 2010. Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing* 19(7):1908–1920.

Xiao, L. 2010. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research* 11:2543–2596.

Yang, L., and Jin, R. 2006. Distance metric learning: A comprehensive survey. classification. *Michigan State University* 1–51.

Yang, L.; Jin, R.; Mummert, L. B.; Sukthankar, R.; Goode, A.; Zheng, B.; Hoi, S. C. H.; and Satyanarayanan, M. 2010. A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(1):30–44.

Zhai, D.; Chang, H.; Shan, S.; Chen, X.; and Gao, W. 2012. Multiview metric learning with global consistency and local smoothness. *ACM TIST* 3(3):53.

Zhang, J.; Marszalek, M.; Lazebnik, S.; and Schmid, C. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision* 73(2):213–238.