

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

8-2017

Sparse online learning of image similarity

Xingyu GAO

Chinese Academy of Sciences

Steven C. H. HOI

Singapore Management University, CHHOI@smu.edu.sg

Yongdong ZHANG

Chinese Academy of Sciences

Jianshe ZHOU

Chinese Academy of Sciences


Ji WAN

Chinese Academy of Sciences

See next page for additional authors

DOI: <https://doi.org/10.1145/3065950>

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), and the [Theory and Algorithms Commons](#)

Citation

GAO, Xingyu; HOI, Steven C. H.; ZHANG, Yongdong; ZHOU, Jianshe; WAN, Ji; CHEN, Zhenyu; LI, Jintao; and ZHU, Jianke. Sparse online learning of image similarity. (2017). *ACM Transactions on Intelligent Systems and Technology*. 8, (5), 64: 1-22. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/3794

This Journal Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Author

Xingyu GAO, Steven C. H. HOI, Yongdong ZHANG, Jianshe ZHOU, Ji WAN, Zhenyu CHEN, Jintao LI, and Jianke ZHU

Sparse Online Learning of Image Similarity

XINGYU GAO, Laboratory of Parallel Software and Computational Science, Institute of Software,
Chinese Academy of Sciences

STEVEN C. H. HOI, School of Information Systems, Singapore Management University

YONGDONG ZHANG, Institute of Computing Technology, Chinese Academy of Sciences

JIAN SHE ZHOU, Beijing Advanced Innovation Center for Imaging Technology, Capital Normal
University

JI WAN, Institute of Computing Technology, Chinese Academy of Sciences

ZHENYU CHEN, China Electric Power Research Institute

JINTAO LI, Institute of Computing Technology, Chinese Academy of Sciences

JIANKE ZHU, College of Computer Science and Technology, Zhejiang University

Learning image similarity plays a critical role in real-world multimedia information retrieval applications, especially in Content-Based Image Retrieval (CBIR) tasks, in which an accurate retrieval of visually similar objects largely relies on an effective image similarity function. Crafting a good similarity function is very challenging because visual contents of images are often represented as feature vectors in high-dimensional spaces, for example, via bag-of-words (BoW) representations, and traditional rigid similarity functions, for example, cosine similarity, are often suboptimal for CBIR tasks. In this article, we address this fundamental problem, that is, learning to optimize image similarity with sparse and high-dimensional representations from large-scale training data, and propose a novel scheme of Sparse Online Learning of Image Similarity (SOLIS). In contrast to many existing image-similarity learning algorithms that are designed to work with low-dimensional data, SOLIS is able to learn image similarity from large-scale image data in sparse and high-dimensional spaces. Our encouraging results showed that the proposed new technique achieves highly competitive accuracy as compared to the state-of-the-art approaches but enjoys significant advantages in computational efficiency, model sparsity, and retrieval scalability, making it more practical for real-world multimedia retrieval applications.

CCS Concepts: • **Theory of computation** → **Online learning algorithms**;

This work was supported in part by the National Key Research and Development Program of China (grant no. 2016YFB0800403), the National Nature Science Foundation of China (grant nos. 61525206, 61572472, 61428207), the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 Grant, the Beijing Natural Science Foundation (grant no. 4152050), and the Beijing Advanced Innovation Center for Imaging Technology (grant no. BAICIT-2016009).

Authors' addresses: X. Gao, Laboratory of Parallel Software and Computational Science, Institute of Software, Chinese Academy of Sciences, Beijing, 100190, China; email: xingyu@iscas.ac.cn; S. C. H. Hoi, School of Information Systems, Singapore Management University, 188065 Singapore; email: chhoi@smu.edu.sg; X. Gao, Y. Zhang, J. Wan, Z. Chen, and J. Li, Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China; emails: {gaoxingyu, zhyd, wanji, chenzhenyu, jtli}@ict.ac.cn; Z. Chen, China Electric Power Research Institute, Beijing, 100192, China; email: chenzhenyu@epri.sgcc.com.cn; J. Zhou, Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing, 100048, China; email: zhouyz0507@icloud.com; J. Zhu, College of Computer Science and Technology, Zhejiang University, Hangzhou, 310058, Zhejiang, China; email: jkzhu@zju.edu.cn.

*Steven C. H. Hoi and Y. Zhang are the corresponding authors.

Additional Key Words and Phrases: Online learning, metric learning, similarity learning, distance metric, bag-of-words representation, image retrieval

1. INTRODUCTION

Recently, there has been an explosive growth of multimedia data on the Internet due to the popularity of social networking and social media applications [Gao et al. 2013a, 2013b]. For many real-world multimedia applications, a fundamental research task is to compute image similarity [Mei and Rui 2009], which 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; the second is to study effective and efficient distance/similarity functions over the features. For feature representation, researchers in multimedia and computer vision have proposed a variety of features for Content-Based Image Retrieval (CBIR) over the past decade [Rahmani et al. 2008]. Examples include global features, color, texture, and shape [Gevers and Smeulders 2000], and local features, SIFT feature descriptors [Lowe 2004; Mikolajczyk and Schmid 2005; Quelhas et al. 2007] and SURF feature descriptors [Bay et al. 2006], as well as their Bag-of-Words (BoW) representations [Fergus et al. 2005; Wang et al. 2006; Bosch et al. 2007; Wu et al. 2010; Jegou et al. 2010]. For distance/similarity functions, a variety of schemes have also been proposed in multimedia and computer vision. The commonly used approaches include Cosine similarity and Euclidean distance, both of which assume a rigid similarity or distance function in some vector space that is often not optimal in real applications [Zheng et al. 2015; Pan et al. 2016b; Xia et al. 2016; Zhili et al. 2016; Zhou et al. 2017].

To overcome the limitations of rigid distance/similarity functions, Distance Metric Learning (DML) techniques [Yang and Jin 2006] have been actively explored to optimize distance metrics, and have been found promising results in various applications, such as image retrieval [Zha et al. 2009; Hoi et al. 2008; Yang et al. 2012; Wu et al. 2013; Zhang et al. 2014; Pan et al. 2014; Wan et al. 2015; Wu et al. 2016], image and video annotation [Mei et al. 2008; Wu et al. 2011; Xu et al. 2011], and mobile application [Chen et al. 2013, 2015; Gao et al. 2016]. Specifically, a typical task of DML is to optimize the generalized Mahalanobis distance of two instances in some vector space as follows:

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

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 (PSD) in order to satisfy the metric property, that is, $\mathbf{M} \geq 0$. Despite being studied extensively, a major drawback of the existing DML schemes is that imposing the PSD constraint often results in computationally intensive algorithms. In addition, many existing DML algorithms usually work in batch-learning mode, which scales poorly for large amounts of training data.

Chechik et al. [2010] had attempted to overcome these limitations by avoiding the PSD constraint when learning the similarity functions in an online learning approach. Specifically, instead of learning the generalized Mahalanobis distance, they proposed OASIS—a novel scheme that attempts to learn the following parametric bilinear similarity function:

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

where $\mathbf{W} \in \mathbb{R}^{m \times m}$ does not need to satisfy the PSD constraint. Although OASIS is much more efficient and scalable than the previous DML algorithms for optimizing the BoW representation with relatively small vocabulary sizes (e.g., 1k–10k), they would still suffer from the critical challenge of poor computational efficiency when handling large-scale, very-high-dimensional BoW data with large vocabulary sizes.

To tackle this challenge, in this article, we propose a novel scheme of Sparse Online Learning of Image Similarity (SOLIS) to effectively exploit the sparsity of visual image feature representations for learning image similarity efficiently from large-scale sparse BoW representations with very large vocabulary sizes. Specifically, the proposed SOLIS scheme attempts to tackle the online image-similarity learning task by exploring the recent advances of sparse online-learning techniques in machine learning [Langford et al. 2009; Xiao 2010]. Unlike the existing OASIS approach, which does not guarantee finding sparse similarity functions, SOLIS is able to learn sparse similarity functions, which enjoys several salient advantages over OASIS when learning to optimize BoW, including (i) *space efficiency*—it significantly reduces storage cost by discarding a large fraction of noninformative codewords with zero weights; (ii) *indexing and retrieval efficiency*—it considerably reduces indexing and retrieval time cost using the resulting compact codebook; and (iii) *learning efficiency*—it can also significantly reduce the computational cost for learning the weights from large-scale training data due to the proposed efficient and scalable sparse online learning of the image-similarity scheme.

In summary, these are the main contributions of this article:

- We present a novel framework of Sparse Online Learning of Image Similarity (SOLIS) for learning sparse similarity functions from large-scale sparse high-dimensional data.
- We propose a family of four different SOLIS 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 methods for optimizing the BoW representation in image retrieval.

We note that a short version of this work has been published in AAAI2014 [Gao et al. 2014]. This new version has been significantly extended and rewritten by including a substantial amount of new content. The rest of this article is organized as follows. Section 2 briefly reviews related work. The system framework of our SOLIS scheme is described in Section 3. Section 4 presents the problem formulation and proposed algorithms with application to image retrieval. Section 5 contains extensive experimental results and discussions. We present our conclusions in Section 6.

2. RELATED WORK

In this section, we review two major categories of related work in multimedia [Chen et al. 2015; Li et al. 2015; Wang et al. 2016], computer vision [Pan et al. 2015, 2016a] and machine learning.

2.1. Distance and Similarity Learning

Our work is closely related to DML [Hoi et al. 2006] or similarity learning [Xia et al. 2014], which has been extensively studied in the literature [Yang and Jin 2006]. A variety of algorithms have been proposed by following different settings and methodologies across different communities. In terms of training data formats, most existing works can be generally grouped into two major categories: (i) learning distance/similarity functions directly from explicit class labels [Weinberger and Saul 2009] that are common for generic data classification tasks and (ii) learning distance/similarity functions from side information [Wu et al. 2009] (either pairwise [Hoi et al. 2008] or triplet constraints [Chechik et al. 2010]), which are common for multimedia retrieval applications.

In terms of learning methodology, most existing methods often adopt batch machine-learning approaches. The major limitation of this learning methodology is that the model has to be retrained from scratch whenever there is new training data. In recent years, some emerging studies have attempted to explore online-learning techniques to tackle the learning tasks [Jain et al. 2009; Chechik et al. 2010] in an efficient and scalable way. Our work also follows the online-learning methodology [Hoi et al. 2014] to tackle image-similarity learning tasks.

Although various techniques have been proposed for learning image distance metrics or similarity functions in the literature, one common issue with the existing approaches is that they often learn a full matrix from relatively low-dimensional image representations (e.g., typical DML studies [Yang and Jin 2006]) or sometimes learn a dense diagonal matrix from high-dimensional BoW representations (e.g., OASIS [Chechik et al. 2010]), either of which often results in computationally intensive solutions, making them hardly scalable for very high-dimensional data. In addition, learning a full matrix for the distance metric or a dense diagonal matrix for similarity functions also will lead to high computational cost when deploying the distance/similarity functions in the final applications. Unlike the existing approaches, our goal is to study a highly efficient and scalable online learning scheme for learning sparse image similarity functions from large-scale very high-dimensional data.

2.2. Sparse Online Learning

Our work is also related to sparse online learning in machine learning [Langford et al. 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 techniques 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 in the literature.

The first group is the family of first-order sparse online-learning algorithms, which 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 has been proposed in Langford et al. [2009], 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 DA 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. Recently, there have been some emerging studies that attempt to explore second-order information for improving sparse online-learning tasks, such as the Adaptive subgradient methods [Duchi et al. 2011], which dynamically exploit knowledge of the geometry of the data observed in previous iterations to perform more informative gradient-based online learning.

Despite being studied actively, most existing works have been focusing on learning classifiers for online classification tasks. For example, Tan et al. [2016] presents the confidence-weighted learning scheme for learning sparse classifiers on high-dimensional data in traditional classification settings, while our work is about learning sparse similarity in retrieval settings. In this work, we apply sparse online-learning techniques for resolving image relative similarity learning tasks, for which the training

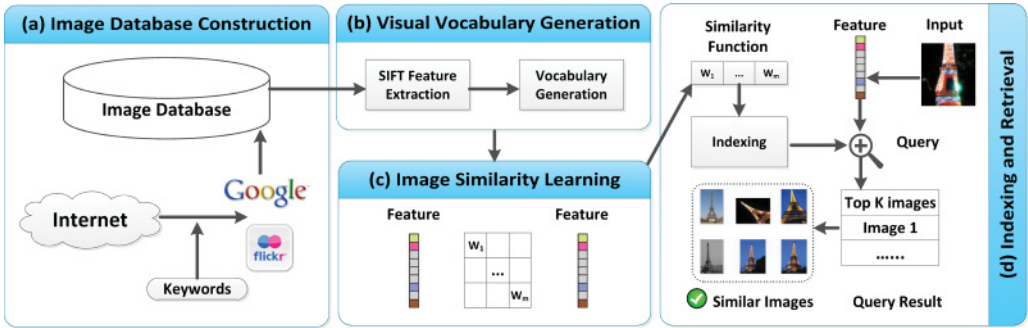


Fig. 1. The system flow of the proposed SOLIS scheme, including (a) Image Database Construction—images collected from regular image benchmark databases or web images crawled from Internet; (b) Visual Vocabulary Generation—extracting SIFT features from images in the databases and then generating visual vocabularies by fast approximate k-means clustering; (c) Image Similarity Learning—learning image similarity functions by the proposed SOLIS algorithms; (d) Indexing and Retrieval—we apply the learned image similarity for indexing the image database and retrieving similar images given a query image.

data is given in the form of triplet constraints. In our framework, we have extensively explored the applications of both first-order and second-order sparse online-learning algorithms to tackle our image-similarity learning problem.

3. SYSTEM OVERVIEW

We first give an overview of the proposed system by exploring sparse online image-similarity learning techniques to optimize the BoW representations in CBIR. Figure 1 illustrates the system flow of our scheme, which consists of the following major stages: (i) image database construction, (ii) visual vocabulary generation, (iii) sparse online learning of image similarity, and (iv) indexing and retrieval. Here, we briefly describe the key idea of each stage.

The first stage is to construct an image database for both retrieval and learning tasks. For example, one can crawl a collection of desired web images from the Internet as the retrieval database. It is also necessary to have a ground truth set to enable the evaluation of image-similarity learning algorithms and comparison of different retrieval schemes.

The second stage is to generate a visual vocabulary, that is, a codebook with a list of visual words. This is typically obtained by extracting local features (e.g., SIFT [Lowe 2004]) from images and then performing fast approximate k-means clustering [Muja and Lowe 2009] to generate a set of clusters as “visual words.” Unlike image-recognition tasks, a large number of clusters is often more preferred in CBIR for achieving high retrieval accuracy. Based on the visual vocabulary, each image can be represented as a sparse feature vector in a high-dimensional space.

The third stage is to learn the optimal codebook weights from training data (pairwise or triplet constraints). In our approach, we propose a novel online machine-learning scheme to learn sparse weights for the codebook as obtained in the previous stage. This is the key stage for optimizing the BoW representation, including both discrimination and compactness.

The last stage is to apply the sparse weight vector for indexing images with the BoW representations to retrieval tasks. It is important to note that the sparse weight vector can save a significant amount of indexing and retrieval costs since we can simply discard the visual words of zero weights, making the BoW representation much more compact and efficient for large-scale CBIR.

4. SPARSE ONLINE LEARNING OF IMAGE SIMILARITY

In this section, we first give a formal formulation of the proposed SOLIS that aims to explore machine-learning techniques to optimize the BoW representations for CBIR tasks. We then explore a family of efficient and scalable sparse online-learning algorithms to resolve the research problem.

4.1. Problem Formulation

We address the fundamental problem of image-similarity learning from side information (e.g., training data in the forms of pairwise or triplet image relationship) to CBIR applications. To formulate the image-similarity learning task, we let $S(\mathbf{x}_i, \mathbf{x}_j)$ denote the similarity function between any two images $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^m$ and assume that a collection of training data instances are given sequentially in the form of triplet instances $\{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-), i = 1, \dots, n\}$, where each triplet instance indicates the relationship between three images, that is, 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 $S(\cdot, \cdot)$ that can produce the similarity values always satisfying the triplet constraints as follows:

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

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

In this article, we aim to explore sparse online-learning techniques for optimizing image-similarity functions in CBIR applications, where images are often represented as a sparse Bag-of-Words (BoW) feature vector in high-dimensional spaces. More specifically, we consider the problem of image-similarity learning for optimizing a parametric bilinear similarity function S defined as follows:

$$S_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbf{x}_i^{\top} \mathbf{W} \mathbf{x}_j, \quad (4)$$

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

Given this similarity function and the constraints in Equation (3), we can formulate the problem of image-similarity learning as a convex optimization task

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

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

$$\mathcal{L}((\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-); \mathbf{W}) = \max(0, 1 - S_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_i^+) + S_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_i^-)). \quad (6)$$

Minimizing this loss is equivalent to minimizing the amount of violation on the constraints defined in Equation (3).

This optimization is a batch-learning formulation with a full matrix \mathbf{W} of space complexity $O(m^2)$, which poses a huge challenge when handling large-scale high-dimensional data. In order to deal with very-high-dimensional image data (e.g., millions of dimensions), we simplify the problem by considering the similarity function defined by a diagonal weight matrix: $\mathbf{W} = \text{diag}(\mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^m$. We can rewrite the loss function \mathcal{L} into

$$\mathcal{L}((\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-); \mathbf{W}) = \max(0, 1 - S_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_i^+) + S_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_i^-)), \quad (7)$$

where $S_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbf{x}_i^{\top} \text{diag}(\mathbf{w}) \mathbf{x}_j$.

Instead of solving the optimization task using regular batch-learning algorithms, we propose exploring online-learning techniques to tackle the learning task for several reasons. First of all, online learning avoids the retraining needed by batch-learning algorithms when there is new training data. We note that this is particularly critical for a real-world CBIR application, since training data is often collected from user-relevance feedback or search query logs, and thus usually arrives in a sequential manner along with the development and deployment of a CBIR system. In addition, online-learning algorithms are often simple in nature and usually more efficient and scalable than batch-learning algorithms for very-large-scale applications. Next, we give the details of our online-learning formulations.

By following typical settings of online learning [Hoi et al. 2014], we assume that a triplet instance $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$ is received at every step $t = 1, \dots, n$. The goal of SOLIS is to sequentially update the metric $M = \text{diag}(\mathbf{w})$ by solving the following online optimization task:

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

where $r(\mathbf{w})$ is a sparsity-promoting regularizer, for example, the ℓ_1 -regularizer $r(\mathbf{w}) = \|\mathbf{w}\|_1$ in our approach. In the following, we present a family of efficient and scalable algorithms to tackle the aforementioned optimization task of sparse online image similarity learning for handling very-high-dimensional BoW data.

4.2. SOLIS-TG: SOLIS Algorithm via Truncated Gradient

We first attempt to solve the SOLIS problem by exploring the TG-based techniques [Langford et al. 2009], which extend the online gradient descent with truncation tricks for achieving sparsity.

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

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

where $\nabla \mathcal{L}_{\mathbf{w}_t}$ is a subgradient 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 guarantee the production of sparse weights at every online learning step.

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

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

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

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

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

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

where \odot denotes an elementwise product of two vectors. This 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 SOLIS-TG algorithm.

4.3. SOLIS-DA: SOLIS Algorithm via Dual Averaging

Our second solution is to explore Nesterov's DA method [Nesterov 2009] and its extensions [Xiao 2010] to tackle the problem of sparse online learning of image similarity,

ALGORITHM 1: SOLIS-TG—SOLIS via Truncated Gradient

Input: Training triplets: $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$, $t = 1, \dots, n$.

Output: The final weight matrix: $\text{diag}(\mathbf{w}_{n+1})$.

```
1: Initialize  $\mathbf{w}_1 = 0$ ;  $\alpha = \eta\lambda$ 
2: for  $t = 1, \dots, n$  do
3:   Receive a triplet instance  $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$ ,
4:   Suffer loss  $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t)$  measured by Equation (7)
5:   if  $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t) > 0$  then
6:      $\mathbf{v} = \mathbf{w}_t - \eta[\mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-)]$ ;
7:     for  $j=1$  to  $m$  do
8:       if  $\mathbf{v}_j \geq 0$  then
9:          $\mathbf{w}_{t+1,j} = \max(0, \mathbf{v}_j - \alpha)$ ;
10:      else
11:         $\mathbf{w}_{t+1,j} = \min(0, \mathbf{v}_j + \alpha)$ ;
12:      end if
13:    end for
14:  end if
15: end for
```

which attempts to exploit all the past subgradients of the loss function and the whole regularization term instead of using 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 this notation, we can derive the closed-form solution of $\mathbf{w}_{t+1} = [\mathbf{w}_{t+1}^{(1)}, \dots, \mathbf{w}_{t+1}^{(m)}]$ for optimizing Equation (13) as

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

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 Learning of Image Similarity via Dual Averaging (SOLIS-DA) algorithm.

4.4. SOLIS-AFB: SOLIS Algorithm via Adaptive FOBOS

The just presented SOLIS algorithm exploits only the first-order information of the weight vector at each online step. To address this limitation, we propose a second-order

ALGORITHM 2: SOLIS-DA—SOLIS via Dual Averaging

Input:

 1: Training triplets: $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$, $t = 1, \dots, n$

 2: Input parameters: $\gamma > 0, \rho \geq 0$
Output: The final weight matrix: $\text{diag}(\mathbf{w}_{n+1})$.

 3: Initialize $\mathbf{w}_1 = \mathbf{0}, \bar{\mathbf{v}}_0 = \mathbf{0}$

 4: **for** $t = 1, \dots, n$ **do**

 5: Receive a triplet instance $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$,

 6: Suffer loss $\mathcal{L}(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-; \mathbf{w}_t)$ measured by Equation (7)

 7: Compute $\bar{\mathbf{v}}_t = \frac{t-1}{t}\bar{\mathbf{v}}_{t-1} + \frac{1}{t}\mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-)$

 8: Compute $\lambda_t = \lambda + \gamma\rho/\sqrt{t}$

 9: **if** $\mathcal{L}(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-; \mathbf{w}_t) > 0$ **then**

 10: **for** $j=1$ to m **do**

 11: **if** $|\bar{v}_t^{(j)}| \leq \lambda_t$ **then**

 12: $\mathbf{w}_{t+1}^{(j)} = 0$;

 13: **else**

 14: $\mathbf{w}_{t+1}^{(j)} = -\frac{\sqrt{t}}{\gamma}(\bar{v}_t^{(j)} - \lambda_t \text{sgn}(\bar{v}_t^{(j)}))$;

 15: **end if**

 16: **end for**

 17: **end if**

 18: **end for**

sparse weight-learning scheme by exploring another state-of-the-art sparse online-learning method, that is, the Adaptive FOBOS method [Duchi et al. 2011], which dynamically exploits knowledge of the geometry of the data observed in previous iterations to perform more informative gradient-based online learning.

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 the composite mirror descent method with ℓ_1 -regularization, as follows:

$$\mathbf{w}_{t+1} \leftarrow \arg \min_{\mathbf{w}} \langle \eta \mathbf{g}_t, \mathbf{w} \rangle + \eta \lambda \|\mathbf{w}\|_1 + B_{\Psi_t}(\mathbf{w}, \mathbf{w}_t), \quad (17)$$

where $B_{\Psi_t}(\mathbf{w}, \mathbf{w}_t) = \Psi_t(\mathbf{w}) - \Psi_t(\mathbf{w}_t) - \langle \nabla \Psi_t(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \rangle$ is the Bregman divergence associated with a strongly convex and differentiable proximal function $\Psi_t(\mathbf{w}) = \frac{1}{2} \langle \mathbf{w}, H_t \mathbf{w} \rangle$. η is the learning rate and $\lambda \geq 0$ is the sparsity-promoting parameter.

$$\begin{aligned} H_t &= \delta I + \text{diag}(G_t)^{1/2} \\ G_t &= \sum_{\tau=1}^t \mathbf{g}_\tau \mathbf{g}_\tau^\top \end{aligned} \quad (18)$$

where $\delta \geq 0$ is the parameter to ensure the positive-definite property of the adaptive weighting matrix. Using this notation, we can derive the closed-form solution of $\mathbf{w}_{t+1} = [\mathbf{w}_{t+1}^{(1)}, \dots, \mathbf{w}_{t+1}^{(m)}]$ for optimizing Equation (17) as

$$\mathbf{w}_{t+1}^{(i)} = \text{sgn} \left(\mathbf{w}_t^{(i)} - \frac{\eta}{H_{t,ii}} \mathbf{g}_t^{(i)} \right) \left[\left| \mathbf{w}_t^{(i)} - \frac{\eta}{H_{t,ii}} \mathbf{g}_t^{(i)} \right| - \frac{\lambda \eta}{H_{t,ii}} \right]_+ \quad (19a)$$

$$\mathbf{g}_t = \mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-) \quad (19b)$$

Algorithm 3 shows the detailed procedures of the proposed Sparse Online Learning of Image Similarity via Adaptive FOBOS (SOLIS-AFB) algorithm.

ALGORITHM 3: SOLIS-AFB—SOLIS via Adaptive FOBOS

Input:

- 1: Training triplets: $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$, $t = 1, \dots, n$
- 2: Input parameters: $\lambda > 0, \eta > 0, \delta \geq 0$

Output: The final weight matrix: $\text{diag}(\mathbf{w}_{n+1})$.

- 3: Initialize $\mathbf{w}_1 = \mathbf{0}$
 - 4: **for** $t = 1, \dots, n$ **do**
 - 5: Receive a triplet instance $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$;
 - 6: Suffer loss $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t)$ measured by Equation (7);
 - 7: **if** $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t) > 0$ **then**
 - 8: Compute $\mathbf{g}_t = \mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-)$;
 - 9: Compute $G_t = \sum_{\tau=1}^t \mathbf{g}_\tau \mathbf{g}_\tau^\top$;
 - 10: Compute $H_t = \delta I + \text{diag}(G_t)^{1/2}$;
 - 11: **for** $j=1$ to m **do**
 - 12: $\mathbf{w}_{t+1}^{(j)} = \text{sgn} \left(\mathbf{w}_t^{(j)} - \frac{\eta}{H_{t,jj}} \mathbf{g}_t^{(j)} \right) \left[\left| \mathbf{w}_t^{(j)} - \frac{\eta}{H_{t,jj}} \mathbf{g}_t^{(j)} \right| - \frac{\lambda \eta}{H_{t,jj}} \right]_+$;
 - 13: **end for**
 - 14: **end if**
 - 15: **end for**
-

4.5. SOLIS-ADA: SOLIS Algorithm via Adaptive RDA

Our fourth solution is to explore the Adaptive Regularized Dual Averaging method [Duchi et al. 2011], which dynamically exploits knowledge of the geometry of the data observed in previous iterations to perform more informative gradient-based online learning.

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 (RDA) method with ℓ_1 -regularization as follows:

$$\mathbf{w}_{t+1} \leftarrow \arg \min_{\mathbf{w}} \langle \eta \bar{\mathbf{g}}_t, \mathbf{w} \rangle + \eta \lambda \|\mathbf{w}\|_1 + \frac{1}{t} \Psi_t(\mathbf{w}), \quad (20)$$

where $\Psi_t(\mathbf{w}) = \frac{1}{2} \langle \mathbf{w}, H_t \mathbf{w} \rangle$ is the same proximal function as in *SOLIS-AFB*. $\bar{\mathbf{g}} = \frac{1}{t} \sum_{\tau=1}^t \mathbf{g}_\tau$ is the average subgradient of \mathcal{L} at the online step. η is the learning rate and $\lambda \geq 0$ is the sparsity-promoting parameter.

$$\begin{aligned} H_t &= \delta I + \text{diag}(G_t)^{1/2} \\ G_t &= \sum_{\tau=1}^t \mathbf{g}_\tau \mathbf{g}_\tau^\top, \end{aligned} \quad (21)$$

where $\delta \geq 0$ is the parameter to ensure that positive-definite property of the adaptive weighting matrix. Using this notation, we can derive the closed-form solution of $\mathbf{w}_{t+1} = [\mathbf{w}_{t+1}^{(1)}, \dots, \mathbf{w}_{t+1}^{(m)}]$ for optimizing Equation (20) as

$$\mathbf{w}_{t+1}^{(i)} = \text{sgn}(-t \bar{\mathbf{g}}_t^{(i)}) \frac{\eta t}{H_{t,ii}} \left[|\bar{\mathbf{g}}_t^{(i)}| - \lambda \right]_+ \quad (22a)$$

$$\mathbf{g}_t = \mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-). \quad (22b)$$

Algorithm 4 shows the detailed procedures of the proposed Sparse Online Learning of Image Similarity via Adaptive RDA (SOLIS-ADA) algorithm.

5. EXPERIMENTS

In our experiments, we investigate the application of the proposed SOLIS technique for improving the BoW representation in CBIR tasks. In the following, we first introduce

ALGORITHM 4: SOLIS-ADA—SOLIS via Adaptive RDA

Input:

- 1: Training triplets: $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$, $t = 1, \dots, n$
- 2: Input parameters: $\lambda > 0$, $\eta > 0$, $\delta \geq 0$

Output: The final weight matrix: $\text{diag}(\mathbf{w}_{n+1})$.

- 3: Initialize $\mathbf{w}_1 = \mathbf{0}$
 - 4: **for** $t = 1, \dots, n$ **do**
 - 5: Receive a triplet instance $(\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-)$;
 - 6: Suffer loss $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t)$ measured by Equation (7);
 - 7: **if** $\mathcal{L}((\mathbf{x}_t, \mathbf{x}_t^+, \mathbf{x}_t^-); \mathbf{w}_t) > 0$ **then**
 - 8: Compute $\mathbf{g}_t = \mathbf{x}_t \odot (\mathbf{x}_t^+ - \mathbf{x}_t^-)$;
 - 9: Compute $\mathbf{G}_t = \sum_{\tau=1}^t \mathbf{g}_\tau \mathbf{g}_\tau^\top$;
 - 10: Compute $H_t = \delta I + \text{diag}(\mathbf{G}_t)^{1/2}$;
 - 11: **for** $j=1$ to m **do**
 - 12: $\mathbf{w}_{t+1}^{(j)} = \text{sgn}(-t \bar{\mathbf{g}}_t^{(j)}) \frac{\eta t}{H_{t,ii}} [|\bar{\mathbf{g}}_t^{(j)}| - \lambda]_+$;
 - 13: **end for**
 - 14: **end if**
 - 15: **end for**
-

the experimental testbed and setup, followed by presenting the detailed experimental results and discussions.

5.1. Experimental Testbed and Setup

We use the BoW representation for representing the images in our datasets. Specifically, we use SIFT descriptors [Lowe 2004] and fast Approximate K-Means (AKM) clustering [Muja and Lowe 2009] to generate codebooks of varying sizes in three different scales: 10K (10,000), 100K (100,000), and 1M (1-million). We conduct performance evaluation on several publicly available image datasets, including *Oxford5K*¹ [Philbin et al. 2007], *Paris*² [Philbin et al. 2008], and 1-million Flickr images *MIRFlickr1M*³ [Huiskes et al. 2010]. More details about these datasets will be discussed in subsequent sections.

In the following experiments, we first evaluate the retrieval quality of different methods measured by mean Average Precision (mAP) followed by evaluating the model sparsity for all cases. Finally, we will also evaluate the computational time costs for training, indexing, and retrieval by different schemes.

5.2. Comparison Algorithms

In order to examine the efficacy of the proposed SOLIS 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 et al. 1999];
- QPAO: A state-of-the-art codebook learning approach [Cai et al. 2010] which formulates it as quadratic programming (QP) and adopts Alternating Optimization (AO) to solve it;
- OASIS: Online Algorithm for Scalable Image Similarity [Chechik et al. 2010], a state-of-the-art online learning algorithm for image-similarity learning;
- LEGO: a state-of-the-art online metric-learning algorithm [Jain et al. 2009] for similarity search;

¹<http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/index.html>.

²<http://www.robots.ox.ac.uk/~vgg/data/parisbuildings/index.html>.

³http://press.liacs.nl/mirflickr/#sec_download.

- LMNN: Largest Margin Nearest Neighbor [Weinberger and Saul 2009], a state-of-the-art DML algorithm in which k nearest neighbors belong to the same class while instances from different classes are separated by the large margin;
- SOLIS: the four proposed SOLIS algorithms, including SOLIS-TG, SOLIS-DA, SOLIS-AFB, and SOLIS-ADA, which are denoted as S-TG, S-DA, S-AFB, and S-ADA, respectively.

For parameter settings, we follow the typical empirical studies of online learning by choosing the best parameters for each algorithm using a separate validation set of randomly sampled triplet sequences for each dataset in our experiments.

5.3. Experiments on Regular Benchmark Datasets

Following previous studies, we adopt the well-known “Oxford5K” image 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, in which 5 images per landmark are used for each query. The 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 compared algorithms on 7 landmarks out of 11 and exclude another 4 landmarks because they simply have too few training examples to learn a good codebook by the algorithms. The same setting was also adopted by the previous codebook learning study in Cai et al. [2010].

5.3.1. Evaluation of Mean Average Precision. Table I shows the evaluation of mAP performance by different schemes. We can draw several observations from the results.

First, we observe that most learning schemes are able to outperform the unsupervised TF-IDF scheme for most cases. This shows the efficacy and importance of optimizing the BoW representation in CBIR by applying machine-learning techniques in exploiting side information/training data.

Second, we found that the existing batch-learning approach QPAO generally achieved better retrieval performance than the two online algorithms OASIS and LEGO. This is perhaps not too surprising because QPAO solves a batch optimization that thus might get better solutions, while all online algorithms (OASIS, LEGO, and our SOLIS algorithms) only learn from a single pass of the triplet instances.

Further, by examining the four proposed SOLIS algorithms (S-TG, S-DA, S-AFB, and S-ADA), we found that their overall retrieval performance is generally much better than OASIS, LEGO, and LMNN, which indicates that the proposed sparsity-inducing image-similarity learning algorithm is potentially more effective than the existing algorithms for online similarity learning without exploiting sparsity. Moreover, by comparing with the batch QPAO algorithm, we found that our SOLIS algorithms are 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 sparsity-inducing online-learning scheme for improving the BoW performance.

Finally, the four proposed SOLIS algorithms achieve very comparable retrieval performance in which the two second-order SOLIS algorithms (S-AFB and S-ADA) tend to slightly outperform the two first-order SOLIS algorithms (S-TG and S-DA).

5.3.2. Evaluation of Sparsity of the Learned Weights. The sparsity of BoW plays a critical role for large-scale CBIR systems, especially for image indexing and retrieval stages. A sparse BoW model not only can speed up the indexing and retrieval processes but also can save a significant amount of storage cost. Later, we measure the sparsity of

Table I. Comparison of Mean Average Precision (%) on *Oxford5K* Dataset with Codebooks of Varying Sizes

Codebook Size	Category	TF-IDF	QPAO	OASIS	LEGO	LMNN	S-TG	S-DA	S-AFB	S-ADA
10,000	all souls	40.60	57.42	52.72	26.07	40.78	56.68	45.50	63.05	63.01
	ashmolean	30.66	30.02	33.03	27.49	30.63	30.79	35.90	33.05	33.06
	bodleian	30.11	68.28	65.13	39.24	33.55	64.66	74.53	62.72	62.57
	christ church	46.35	45.79	43.41	43.66	46.51	53.43	52.42	56.25	56.25
	hertford	31.16	51.47	42.18	28.66	31.20	44.30	35.93	47.37	47.28
	magdalen	5.92	8.86	9.34	3.95	5.97	12.12	17.55	15.59	15.41
	radcliffe camera	52.22	82.44	75.12	68.43	53.30	80.68	74.71	78.03	78.07
	mAP	33.86	49.18	45.85	33.93	34.56	48.95	48.08	50.87	50.81
100,000	all souls	58.17	93.92	75.22	58.25	58.29	91.58	88.70	91.33	91.09
	ashmolean	44.96	42.78	47.68	43.96	44.96	40.15	41.12	44.35	44.11
	bodleian	49.06	86.02	71.36	60.02	52.66	83.07	83.42	85.58	85.67
	christ church	52.08	70.74	50.97	52.06	52.10	59.73	59.04	66.10	65.95
	hertford	53.51	63.93	57.75	52.81	53.51	63.41	60.44	76.60	76.61
	magdalen	11.29	10.99	12.96	6.91	11.29	9.63	18.42	19.50	19.95
	radcliffe camera	70.51	82.19	76.92	70.34	70.51	76.69	76.43	85.16	85.19
	mAP	48.51	64.37	56.13	49.19	49.05	60.61	61.08	66.95	66.94
1,000,000	all souls	53.96	62.99	55.03	53.96	*	62.71	63.57	63.98	64.03
	ashmolean	53.86	48.77	53.45	53.86	*	48.63	51.53	53.14	53.40
	bodleian	66.88	90.21	70.57	66.88	*	84.17	83.47	91.48	91.80
	christ church	56.67	65.36	57.34	56.66	*	61.10	61.10	62.24	62.22
	hertford	72.00	68.66	75.36	72.00	*	79.24	82.78	83.83	83.83
	magdalen	18.98	15.63	19.24	19.01	*	8.79	9.77	9.58	9.58
	radcliffe camera	64.43	62.94	65.04	64.42	*	58.03	61.85	62.55	62.38
	mAP	55.25	59.23	56.58	55.26	*	57.53	59.15	60.97	61.04

Note: In the results, “*” denotes the case in which a method cannot be completed within 5 days.

the learned weights by different algorithms, that is, the number of zero values in the learned weight vectors.

Table II shows the sparsity evaluation of the learned weights by different learning approaches with 10K-sized, 100K-sized, and 1M-sized codebooks. We can draw several observations from the results that follow.

First, we found that OASIS, LEGO and LMNN fail to produce sparse weights for most cases, especially for large-sized codebooks. 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 four proposed SOLIS algorithms are able to produce sparse weights for all cases. In particular, it seems that the larger the codebook size, the higher the sparsity achieved by the proposed algorithms. Finally, by comparing the four proposed SOLIS algorithms themselves, the second-order SOLIS algorithms (S-AFB and S-ADA) generally achieves better sparsity than the first-order algorithms (S-TG and S-DA) for most cases primarily because it exploits all past subgradients and thus achieves better sparsity.

5.3.3. Evaluation of Sparsity versus mAP. Figure 2 shows the sparsity versus mAP on the category of *hertford* in the *Oxford5K* dataset by different approaches with codebooks of varying sizes. From the empirical results, we found that both LEGO and LMNN fail to achieve higher mAP and produce sparse weights for most cases, especially for large-sized codebooks. Similarly, QPAO is able to produce reasonably sparse weights on the 10K-sized codebook with better performance but also fails when the codebook sizes are too large. By contrast, the four proposed SOLIS algorithms can keep very good mAP performances even when the sparsity is very high. Moreover, we observe that the

Table II. Comparison of Sparsity Rate (%) of Learned Weights by Different Approaches on *Oxford5K* Dataset with Codebooks of Varying Sizes

Codebook Size	Category	QPAO	OASIS	LEGO	LMNN	S-TG	S-DA	S-AFB	S-ADA
10,000	all souls	44.99	0.00	0.00	0.00	22.36	64.54	86.73	88.89
	ashmolean	40.16	0.00	0.00	0.00	25.10	74.49	91.59	93.24
	bodleian	35.80	0.00	0.00	0.00	77.43	93.43	96.01	96.35
	christ church	31.91	0.00	0.00	0.00	32.08	76.92	93.36	94.62
	hertford	43.19	0.00	0.00	0.00	22.87	61.82	85.99	88.11
	magdalen	47.59	0.00	0.00	0.00	19.62	52.61	85.83	88.72
	radcliffe camera	43.37	0.00	0.00	0.00	40.48	75.53	89.64	91.09
100,000	all souls	0.02	0.00	0.00	0.00	90.64	97.82	98.50	98.69
	ashmolean	0.02	0.00	0.00	0.00	86.37	98.11	98.83	98.99
	bodleian	0.02	0.00	0.00	0.00	95.25	98.82	99.45	99.51
	christ church	0.01	0.00	0.00	0.00	91.42	98.87	99.17	99.26
	hertford	0.00	0.00	0.00	0.00	92.99	97.56	98.27	98.47
	magdalen	0.04	0.00	0.00	0.00	82.26	97.04	98.22	98.45
	radcliffe camera	0.01	0.00	0.00	0.00	93.47	97.18	98.16	98.41
1,000,000	all souls	0.00	0.00	0.00	*	99.51	99.88	99.93	99.93
	ashmolean	0.00	0.00	0.00	*	99.30	99.96	99.97	99.97
	bodleian	0.00	0.00	0.00	*	99.65	99.85	99.90	99.91
	christ church	0.00	0.00	0.00	*	99.67	99.97	99.97	99.97
	hertford	0.00	0.00	0.00	*	99.77	99.90	99.90	99.90
	magdalen	0.00	0.00	0.00	*	99.06	99.96	99.96	99.96
	radcliffe camera	0.00	0.00	0.00	*	99.51	99.82	99.84	99.88

Note: In the results, “*” denotes cases in which a method cannot be completed within 5 days.

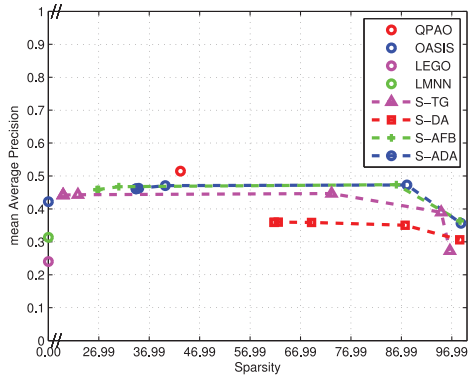
larger the codebook size, the higher the sparsity achieved by the proposed algorithms without slashing mAP performance. Finally, by comparing the four SOLIS algorithms, the second-order algorithms (S-AFB and S-ADA) generally outperform the first-order algorithms (S-TG and S-DA) for varied-sparsity cases.

5.3.4. Evaluation of Computational Cost. Finally, we evaluate the computational costs of different schemes.

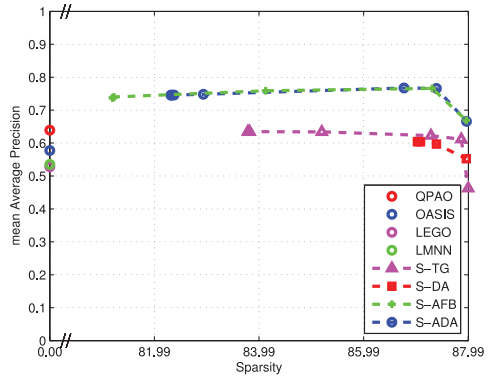
Training Time Cost. Table III shows the evaluation results of training time costs by different schemes on the *Oxford5K* dataset with three codebooks of varying sizes.

We can draw some observations from the results. First, we can see that QPAO and LMNN are the least efficient algorithms. 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 and LEGO are far more efficient than QPAO and LMNN on relatively lower-dimensional space since OASIS and LEGO are online algorithms with time complexity linear with respect to the sample size and dimensionality. However, when handling very-high-dimensional data (e.g., 1-million-sized codebook), OASIS and LEGO become inefficient as the dimensionality plays a dominating factor.

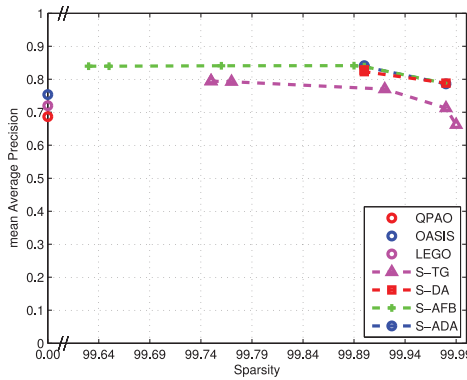
By contrast, the proposed SOLIS algorithms are far more efficient and scalable than all the existing algorithms. Finally, unlike the other algorithms, it is interesting to observe that increasing the dimensionality does not increase the time cost of the proposed algorithms. This seems a bit counterintuitive, but is not difficult to explain. This is because our algorithms always learn sparse weights in the online learning process, and thus the time complexity of our algorithm depends on the number of non-zero elements in the training data instead of the dimensionality. This encouraging



(a) 10K



(b) 100K



(c) 1M

Fig. 2. Sparsity versus mAP on category *hertford* of *Oxford5K* dataset by different approaches with codebooks of varying sizes.

result again validates the efficiency and advantage of the proposed SOLIS technique for large-scale applications.

Indexing Time Cost. To further examine how the sparse BoW model can benefit index construction in the visual similarity search task of CBIR, we evaluate the computational costs of building an index using inverted index techniques [Sivic and Zisserman 2003] for image-retrieval tasks. Figure 3 shows the experimental results of average time cost for indexing the images on the category of *hertford* in the *Oxford5K* dataset by different schemes with codebooks of varying sizes. From the results, we can see that SOLIS can significantly reduce the indexing time cost by the sparsity-inducing learning approach.

Retrieval Time Cost. To examine if the sparse BoW model can further benefit the subsequent image-similarity search task, we evaluate the computational cost of measuring similarity values in the image-retrieval tasks. Figure 4 shows the experimental results of retrieval time cost over 5 queries on the *hertford* category in the *Oxford5K* dataset by different schemes with codebooks of varying sizes.

As observed from the experimental results, by comparing the BoW model with the TF-IDF, QPAO, OASIS, LEGO, and LMNN methods, our proposed SOLIS algorithms achieve the lowest computational cost for computing similarity in the retrieval phase.

Table III. Evaluation of Training Time Cost (Seconds) by Different Schemes on *Oxford5K* Dataset with Varied-sized Codebooks

Codebook Size	Category	QPAO	OASIS	LEGO	LMNN	S-TG	S-DA	S-AFB	S-ADA
10,000	all souls	2.98×10^3	17.91	4.40	5.84×10^3	8.83	7.85	8.56	8.75
	ashmolean	2.51×10^3	17.34	4.36	5.80×10^3	6.61	6.46	6.87	6.03
	bodleian	2.94×10^3	14.64	3.56	5.76×10^3	8.51	8.92	10.22	9.50
	christ church	1.89×10^3	13.77	4.31	5.87×10^3	3.71	3.56	4.72	4.57
	hertford	2.74×10^3	18.27	4.42	5.97×10^3	6.95	5.58	8.05	7.13
	magdalen	2.49×10^3	17.75	4.33	5.84×10^3	6.23	5.69	7.23	7.78
	radcliffe camera	2.93×10^3	16.82	4.36	5.84×10^3	7.09	7.69	9.56	8.73
100,000	all souls	3.03×10^3	74.01	26.87	6.82×10^4	2.38	1.76	2.22	2.17
	ashmolean	2.63×10^3	73.03	26.56	6.05×10^4	1.64	1.14	1.53	1.46
	bodleian	3.89×10^3	68.48	27.00	5.65×10^4	3.57	2.92	3.61	1.76
	christ church	1.95×10^3	69.08	25.97	5.97×10^4	0.99	0.92	0.93	1.06
	hertford	2.87×10^3	73.78	26.34	6.91×10^4	2.10	1.54	1.94	1.76
	magdalen	3.05×10^3	74.59	26.67	5.98×10^4	1.91	1.26	1.76	1.75
	radcliffe camera	2.97×10^3	75.45	26.50	6.04×10^4	2.68	1.81	2.61	2.26
1,000,000	all souls	2.38×10^4	1.26×10^3	5.56×10^2	*	1.34	1.40	1.47	1.45
	ashmolean	2.44×10^4	1.25×10^3	5.63×10^2	*	1.14	1.01	1.03	1.02
	bodleian	1.94×10^4	1.24×10^3	5.80×10^2	*	2.24	2.02	2.26	2.16
	christ church	2.17×10^4	1.25×10^3	5.71×10^2	*	0.99	0.94	1.02	0.98
	hertford	2.19×10^4	1.26×10^3	5.72×10^2	*	1.65	1.72	1.69	1.68
	magdalen	2.28×10^4	1.26×10^3	5.74×10^2	*	1.05	1.09	1.06	1.09
	radcliffe camera	2.22×10^4	1.26×10^3	5.62×10^2	*	1.65	1.72	1.83	1.79

Note: In the results, “*” denotes cases in which a method cannot be completed within 5 days.

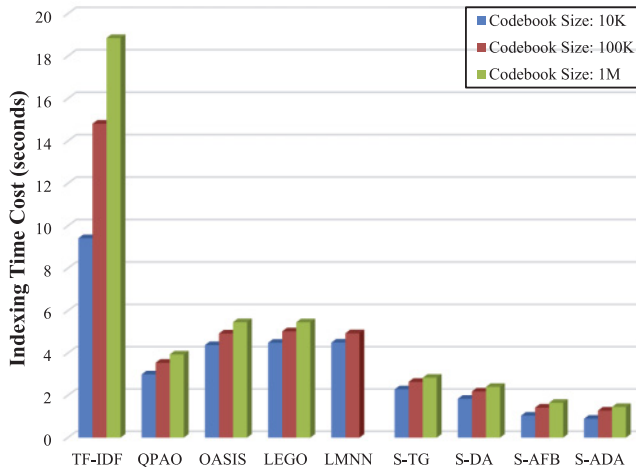


Fig. 3. Comparison of the indexing time cost (seconds) on the category *hertford* of the *Oxford5K* dataset by different schemes with codebooks of varying sizes.

This again validates that the proposed SOLIS scheme can significantly improve the computational efficiency and scalability of image-retrieval tasks.

5.4. Experiments on the Large-Scale Dataset

To further examine the learning efficiency and scalability of the proposed technique, we construct a large-scale image dataset called “*Paris+MIRFlickr1M*” from two public image datasets *Paris* [Philbin et al. 2008] and *MIRFlickr1M* [Huiskes et al. 2010]. The

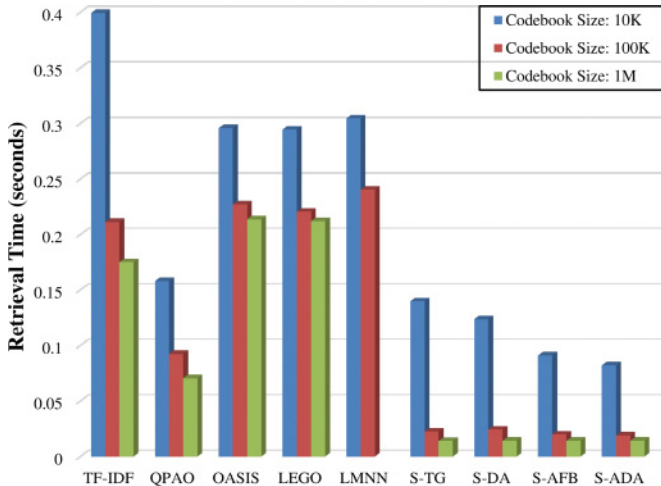


Fig. 4. Comparison of the retrieval time (seconds) on the category *herford* of the *Oxford5K* dataset by different schemes with codebooks of varying sizes.

Table IV. Comparison of Mean Average Precision (%) on *Paris+MIRFlickr1M* Dataset with Codebooks of Varying Sizes

Codebook Size	TF-IDF	OASIS	LEGO	S-TG	S-DA	S-AFB	S-ADA
10K	8.72	11.98	9.24	10.58	11.27	11.35	12.06
100K	26.94	31.33	27.23	28.08	31.95	32.07	32.06
1M	26.52	26.71	26.56	20.57	26.04	26.87	27.17

Paris+MIRFlickr1M dataset consists of images from the *Paris* dataset as ground truth and 1-million web images from *MIRFlickr1M* as distractors, which were downloaded from Flickr [Huiskes et al. 2010]. To generate the triplet instances for our learning task, we repeatedly sample two different images from the same class and another image from a different class randomly according to their ground-truth class labels. Specifically, we generate about 1-million triplet instances as training samples from the *Paris+MIRFlickr1M* dataset with 10K, 100K, and 1M-sized codebooks, respectively.

We first evaluate the retrieval quality of different schemes using the mAP and then evaluate the model sparsity as well as computational efficiency. Finally, we evaluate the time cost for training, indexing, and retrieval by different algorithms with codebooks of varying sizes.

5.4.1. Evaluation of Mean Average Precision. Following a similar protocol, we evaluate the retrieval accuracy of different schemes on the large-scale dataset. Since QPAO and LMNN are very time-consuming and non-scalable for large-scale datasets, we failed to run them successfully on this large dataset. As a result, we can only compare the four proposed SOLIS algorithms with the TF-IDF, OASIS, and LEGO algorithms. Table IV shows the mAP evaluation results on the large-scale *Paris+MIRFlickr1M* dataset.

From the results, we observe that the retrieval performance of the BoW model using TF-IDF decreases considerably on this large-scale image dataset in comparison to the previous experiments, mainly due to the added noisy distracting images. By contrast, the proposed SOLIS schemes still maintain rather high retrieval accuracy. This encouraging result shows that the proposed SOLIS schemes are fairly robust to noisy background images and are able to significantly improve the retrieval accuracy of BoW representations for large-scale complex image-retrieval tasks, in which noise could be quite common.

Table V. Comparison of Model Sparsity (%) on *Paris+MIRFlickr1M* Dataset with Codebooks of Varying Sizes

Codebook Size	OASIS	LEGO	S-TG	S-DA	S-AFB	S-ADA
10K	0.00	0.00	7.12	22.95	91.57	91.97
100K	0.00	0.00	84.11	86.90	99.08	99.13
1M	0.00	0.00	98.62	99.07	99.90	99.08

Table VI. Comparison of Training Time (Seconds) on *Paris+MIRFlickr1M* Dataset with Codebooks of Varying Sizes

Codebook Size	OASIS	LEGO	S-TG	S-DA	S-AFB	S-ADA
10K	1.42×10^2	5.51×10^2	85.76	84.15	82.76	81.05
100K	6.94×10^2	1.97×10^3	75.49	72.96	71.49	70.86
1M	1.24×10^4	1.48×10^4	70.55	68.82	67.55	65.98

5.4.2. Evaluation of Sparsity of the Learned Weights. Similar to the previous experiments, we also conduct experiments to evaluate the sparsity of the BoW models learned by different learning algorithms, which is particularly important for large-scale image-retrieval tasks. Table V shows the experimental results of measuring the sparsity of the BoW weights learned by different algorithms. Similar to the previous results, OASIS and LEGO failed to produce sparse BoW weights, while the four SOLIS algorithms are able to produce the codebook weights with much higher sparsity for all cases, which is critical to speeding up the retrieval process and saving the huge storage cost for large-scale CBIR applications.

5.4.3. Evaluation of Computational Cost. Our last experiment evaluates computational costs of training, indexing, and retrieval by different algorithms with codebooks of varying sizes.

Training Time Cost. Similar to the previous experiments, Table VI shows the comparisons of training time costs by OASIS, LEGO, S-TG, S-DA, S-AFB, and S-ADA on codebooks of varying sizes. From the results, we can see that the proposed SOLIS scheme can handle very-high-dimensional data (e.g., 1M-scale) much more efficiently and scalably than the other approaches. This is because our optimization scheme learns sparse codebook weights through the highly efficient and scalable sparse online-learning technique. This result again validates the significant advantage in learning efficiency for the proposed learning scheme for large-scale applications.

Indexing Time Cost. In order to examine how the sparse BoW model can benefit the indexing task, Figure 5 evaluates the indexing time cost on the *Paris+MIRFlickr1M* dataset with codebooks of varying sizes. Similar observations show that the proposed SOLIS scheme significantly improves the indexing time cost of BoW in CBIR.

Retrieval Time Cost. Finally, to evaluate how the sparse model improves image-searching efficiency, we measure the retrieval time costs for querying the large-scale image dataset *Paris+MIRFlickr1M*. Figure 6 shows the total retrieval time cost using the indexes with codebooks of varying sizes learned by different schemes. The proposed SOLIS algorithms are an order of magnitude faster than the existing algorithms, particularly for large vocabulary sizes, due to the sparsity-inducing advantage. The encouraging results again validate that the proposed SOLIS scheme makes the BoW scheme more practical for large-scale CBIR applications.

5.4.4. Summary and Recommendation. Our previous experiments have examined different aspects of the four proposed SOLIS algorithms. In this section, we give the summary of our empirical observations and our overall recommendation. In particular, our empirical results show that (i) in terms of accuracy, the two second-order sparse

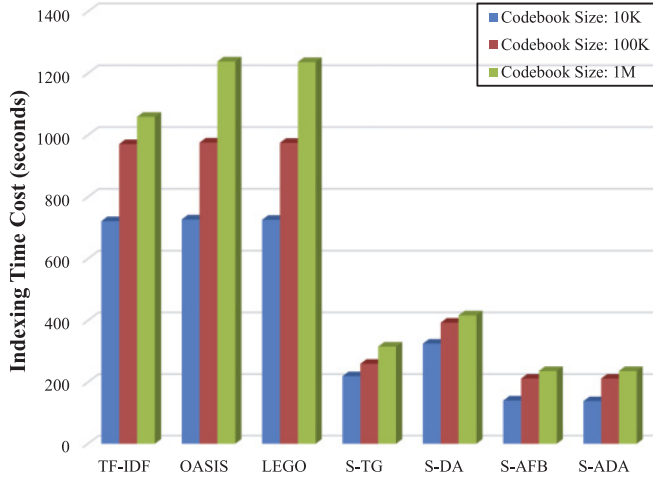


Fig. 5. Comparison of the indexing time cost (seconds) on the *Paris+MIRFlickr1M* dataset with codebooks of varying sizes learned by different schemes.

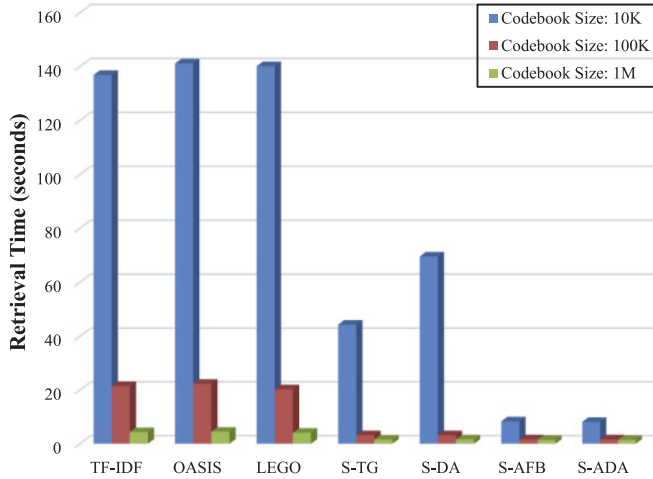


Fig. 6. Comparison of the retrieval time cost (seconds) on the *Paris+MIRFlickr1M* dataset with codebooks of varying sizes learned by different schemes.

online-learning algorithms generally outperform the two first-order learning algorithms in most cases, and there is no significant difference between the two second-order learning algorithms; (ii) in terms of the sparsity of the weights obtained by different algorithms, the two second-order algorithms also outperform the two first-order algorithms in most cases, SOLIS-TG is often the worst, and SOLIS-ADA is the best in most cases; and (iii) finally, in terms of computational cost, we have similar observations in that the two second-order algorithms are slightly more efficient to train and generally faster for retrieval due to the gains of better sparsity. In summary, among the four proposed algorithms, SOLIS-ADA would be the recommended best choice according to our empirical observations.

6. CONCLUSIONS

This article presented SOLIS, the novel scheme Sparse Online Learning of Image Similarity, which aims to optimize Bag-of-Words (BoW) representations by learning from large-scale image data with very-high-dimensional BoW representations. SOLIS explored the recent advances of sparse online learning for tackling the challenging image-similarity learning task, and presented four specific algorithms based on two types of optimization techniques: (i) first-order sparse online learning, that is, TG-based and DA-based learning algorithms; and (ii) second-order sparse online learning, that is, adaptive FOBOS-based and adaptive DA-based learning algorithms. We investigated the application of SOLIS for optimizing the sparse and high-dimensional BoW representations in large-scale CBIR tasks. Our extensive experimental results show that the first-order SOLIS algorithms (SOLIS-TG and SOLIS-DA) and second-order SOLIS algorithms (SOLIS-AFB and SOLIS-ADA) can achieve better or at least comparable retrieval performance than the state-of-the-art approaches but significantly improve both model sparsity and computational cost in training, indexing, and retrieval stages. From the encouraging empirical results, we can conclude that the proposed SOLIS scheme is more effective and promising than the state-of-the-art approaches for optimizing the BoW representations in large-scale CBIR tasks. Among all four variants of the proposed SOLIS algorithms, the SOLIS-ADA using the Adaptive Dual Averaging approach is the overall best recommended choice according to our empirical observations. Finally, we note that the proposed SOLIS technique is not restricted to optimizing BoW representations. In our future work, we plan to explore other types of sparse feature representations using more advanced feature representation learning techniques and apply our technique for building real-world large-scale CBIR systems.

REFERENCES

- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, and others. 1999. *Modern Information Retrieval*. Vol. 463. ACM Press, New York, NY.
- Herbert Bay, Tinne Tuytelaars, and Luc J. Van Gool. 2006. SURF: Speeded up robust features. In *ECCV (1)*. 404–417.
- Anna Bosch, Xavier Muñoz, and Robert Martí. 2007. Which is the best way to organize/classify images by content? *Image and Vision Computing* 25, 6, 778–791.
- Hongping Cai, Fei Yan, and Krystian Mikolajczyk. 2010. Learning weights for codebook in image classification and retrieval. In *CVPR*. 2320–2327.
- Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research* 11, 1109–1135.
- Beijing Chen, Huazhong Shu, Gouenou Coatrieux, Gang Chen, Xingming Sun, and Jean Louis Coatrieux. 2015. Color image analysis by quaternion-type moments. *Journal of Mathematical Imaging and Vision* 51, 1, 124–144.
- Zhenyu Chen, Yiqiang Chen, Xingyu Gao, Shuangquan Wang, Lisha Hu, Chenggang Clarence Yan, Nicholas D. Lane, and Chunyan Miao. 2015. Unobtrusive sensing incremental social contexts using fuzzy class incremental learning. In *IEEE International Conference on Data Mining (ICDM'15)*. IEEE, 71–80.
- Zhenyu Chen, Yiqiang Chen, Shuangquan Wang, Junfa Liu, Xingyu Gao, and Andrew T. Campbell. 2013. Inferring social contextual behavior from Bluetooth traces. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*. ACM, 267–270.
- John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12, 2121–2159.
- John Duchi and Yoram Singer. 2009. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research* 10, 2899–2934.
- Robert Fergus, Fei-Fei Li, Pietro Perona, and Andrew Zisserman. 2005. Learning object categories from Google's image search. In *ICCV*. 1816–1823.
- Xingyu Gao, Juan Cao, Qin He, and Jintao Li. 2013a. A novel method for geographical social event detection in social media. In *Proceedings of the 5th International Conference on Internet Multimedia Computing and Service*. ACM, 305–308.

- Xingyu Gao, Juan Cao, Zhiwei Jin, Xin Li, and Jintao Li. 2013b. GeSoDeck: A geo-social event detection and tracking system. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 471–472.
- Xingyu Gao, Zhenyu Chen, Sheng Tang, Yongdong Zhang, and Jintao Li. 2016. Adaptive weighted imbalance learning with application to abnormal activity recognition. *Neurocomputing* 173, 1927–1935.
- Xingyu Gao, Steven C. H. Hoi, Yongdong Zhang, Ji Wan, and Jintao Li. 2014. SOML: Sparse online metric learning with application to image retrieval. In *AAAI*. 1206–1212.
- Theo Gevers and Arnold W. M. Smeulders. 2000. Pictoseek: Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing* 9, 1, 102–119.
- Steven C. H. Hoi, Wei Liu, Michael R. Lyu, and Wei-Ying Ma. 2006. Learning distance metrics with contextual constraints for image retrieval. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 2072–2078.
- Steven C. H. Hoi, Jialei Wang, and Peilin Zhao. 2014. LIBOL: A library for online learning algorithms. *Journal of Machine Learning Research* 15, 495–499. <http://LIBOL.stevenhoi.org/>.
- Steven C. H. Hoi, Wei Liu, and Shih-Fu Chang. 2008. Semi-supervised distance metric learning for Collaborative Image Retrieval. In *CVPR*.
- Mark J. Huiskes, B. Thomee, and Michael S. Lew. 2010. New trends and ideas in visual concept detection: The MIR Flickr retrieval evaluation initiative. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval (MIR'10)*. ACM, New York, NY, 527–536.
- Prateek Jain, Brian Kulis, Inderjit S. Dhillon, and Kristen Grauman. 2009. Online metric learning and fast similarity search. In *Advances in Neural Information Processing Systems*. 761–768.
- Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Improving bag-of-features for large scale image search. *International Journal of Computer Vision* 87, 3, 316–336.
- John Langford, Lihong Li, and Tong Zhang. 2009. Sparse online learning via truncated gradient. *Journal of Machine Learning Research* 10, 777–801.
- Jian Li, Xiaolong Li, Bin Yang, and Xingming Sun. 2015. Segmentation-based image copy-move forgery detection scheme. *IEEE Transactions on Information Forensics and Security* 10, 3, 507–518.
- David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2, 91–110.
- Tao Mei and Yong Rui. 2009. Image similarity. In *Encyclopedia of Database Systems*. Springer, 1379–1384.
- Tao Mei, Yong Wang, Xian-Sheng Hua, Shaogang Gong, and Shipeng Li. 2008. Coherent image annotation by learning semantic distance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. IEEE, 1–8.
- Krystian Mikolajczyk and Cordelia Schmid. 2005. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 10, 1615–1630.
- Marius Muja and David G. Lowe. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP (1)*. 331–340.
- Yurii Nesterov. 2009. Primal-dual subgradient methods for convex problems. *Mathematical Programming* 120, 1, 221–259.
- Yingwei Pan, Ting Yao, Tao Mei, Houqiang Li, Chong-Wah Ngo, and Yong Rui. 2014. Click-through-based cross-view learning for image search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 717–726.
- Zhaoqing Pan, Peng Jin, Jianjun Lei, Yun Zhang, Xingming Sun, and Sam Kwong. 2016a. Fast reference frame selection based on content similarity for low complexity HEVC encoder. *Journal of Visual Communication and Image Representation* 40, 516–524.
- Zhaoqing Pan, Jianjun Lei, Yun Zhang, Xingming Sun, and Sam Kwong. 2016b. Fast motion estimation based on content property for low-complexity H.265/HEVC encoder. *IEEE Transactions on Broadcasting* 62, 3, 675–684.
- Zhaoqing Pan, Yun Zhang, and Sam Kwong. 2015. Efficient motion and disparity estimation optimization for low complexity multiview video coding. *IEEE Transactions on Broadcasting* 61, 2, 166–176.
- James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*.
- James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*.
- Pedro Quelhas, Florent Monay, Jean-Marc Odobez, Daniel Gatica-Perez, and Tinne Tuytelaars. 2007. A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 9, 1575–1589.

- Rouhollah Rahmani, Sally A. Goldman, Hui Zhang, Sharath R. Cholleti, and Jason E. Fritts. 2008. Localized content-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 11, 1902–1912.
- Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the 9th IEEE International Conference on Computer Vision*. IEEE, 1470–1477.
- Mingkui Tan, Yan Yan, Li Wang, Anton Van Den Hengel, Ivor Tsang, and Qinfeng Javen Shi. 2016. Learning sparse confidence-weighted classifier on very high dimensional data. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- J. Wan, S. Tang, Y. Zhang, et al. 2015. HDIdx: High-dimensional indexing for efficient approximate nearest neighbor search. *Neurocomputing* 237 (2015), 401–404.
- Gang Wang, Ye Zhang, and Fei-Fei Li. 2006. Using dependent regions for object categorization in a generative framework. In *CVPR (2)*. 1597–1604.
- Jinwei Wang, Ting Li, Yun-Qing Shi, Shiguo Lian, and Jingyu Ye. 2016. Forensics feature analysis in quaternion wavelet domain for distinguishing photographic images and computer graphics. *Multimedia Tools and Applications* 1–17.
- Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10, 207–244.
- Lei Wu, Steven C. H. Hoi, Rong Jin, Jianke Zhu, and Nenghai Yu. 2009. Distance metric learning from uncertain side information with application to automated photo tagging. In *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 135–144.
- Lei Wu, Steven C. H. Hoi, and Nenghai Yu. 2010. Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing* 19, 7, 1908–1920.
- Pengcheng Wu, Steven C. H. Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. 2013. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 153–162.
- Pengcheng Wu, Steven C. H. Hoi, Peilin Zhao, Chunyan Miao, and Zhi-Yong Liu. 2016. Online multi-modal distance metric learning with application to image retrieval. *IEEE Transactions on Knowledge and Data Engineering* 28, 2, 454–467.
- Pengcheng Wu, Steven C. H. Hoi, Peilin Zhao, and Ying He. 2011. Mining social images with distance metric learning for automated image tagging. In *WSDM*. 197–206.
- Hao Xia, Steven C. H. Hoi, Rong Jin, and Peilin Zhao. 2014. Online multiple kernel similarity learning for visual search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3, 536–549.
- Zhihua Xia, Xinhui Wang, Liangao Zhang, Zhan Qin, Xingming Sun, and Kui Ren. 2016. A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing. *IEEE Transactions on Information Forensics and Security* 11, 11, 2594–2608.
- Lin Xiao. 2010. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research* 11, 2543–2596.
- Ye Xu, Wei Ping, and Andrew T. Campbell. 2011. Multi-instance metric learning. In *ICDM*. 874–883.
- L. Yang and R. Jin. 2006. Distance metric learning: A comprehensive survey classification. *Michigan State University* 1–51.
- Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, and Yunhe Pan. 2012. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 4, 723–742.
- Zheng-Jun Zha, Tao Mei, Meng Wang, Zengfu Wang, and Xian-Sheng Hua. 2009. Robust distance metric learning with auxiliary knowledge. In *IJCAI*. 1327–1332.
- Hong Zhang, Junsong Yuan, Xingyu Gao, and Zhenyu Chen. 2014. Boosting cross-media retrieval via visual-auditory feature analysis and relevance feedback. In *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 953–956.
- Yuhui Zheng, Byeungwoo Jeon, Danhua Xu, Q. M. Wu, and Hui Zhang. 2015. Image segmentation by generalized hierarchical fuzzy C-means algorithm. *Journal of Intelligent and Fuzzy Systems* 28, 2, 961–973.
- Zhili Zhou, Yunlong Wang, Q. M. Jonathan Wu, Ching-Nung Yang, and Xingming Sun. 2017. Effective and efficient global context verification for image copy detection. *IEEE Transactions on Information Forensics and Security* 12, 1, 48–63.
- Zhili Zhou, Ching-Nung Yang, Xingming Sun, Qi Liu, and Q. M. Jonathan Wu. 2016. Effective and efficient image copy detection with resistance to arbitrary rotation. *IEICE Transactions on Information and Systems* 99, 6, 1531–1540.