Learning Regularized, Query-dependent Bilinear Similarities for Large Scale Image Retrieval

Zhanghui Kuang[1]* Jian Sun[2] Kwan-Yee K. Wong[1]
[1]The University of Hong Kong [2]Microsoft Research Asia

Abstract

An effective way to improve the quality of image retrieval is by employing a query-dependent similarity measure. However, implementing this in a large scale system is non-trivial because we want neither hurting the efficiency nor relying on too many training samples. In this paper, we introduce a query-dependent bilinear similarity measure to address the first issue. Based on our bilinear similarity model, query adaptation can be achieved by simply applying any existing efficient indexing/retrieval method to a transformed version (surrogate) of a query. To address the issue of limited training samples, we further propose a novel angular regularization constraint for learning the similarity measure. The learning is formulated as a Quadratic Programming (QP) problem and can be solved efficiently by a SMO-type algorithm. Experiments on two public datasets and our 1-million web-image dataset validate that our proposed method can consistently bring improvements and the whole solution is practical in large scale applications.

1. Introduction

The goal of content-based image retrieval [26, 20, 24, 16] is to retrieve images from a database that are *semantically similar* to a given query image under some defined similarity measures[1]. In large scale image retrieval systems, a single similarity measure is often defined globally to allow the deployment of an efficient indexing data structure (*e.g.*, inverted file [26, 20], hashing [10, 31], or hierarchical search [16, 25, 30]). However, due to the diversity of queries, a single global similarity measure is often insufficient to produce satisfactory results. For instance, a HoG-like feature [24] is more preferable when a query image has rich structures (*e.g.*, Fig. 1 (a)), whereas a query image with salient spatial color layout (*e.g.*, Fig. 1 (b)) favors a color-based feature [23]. It is therefore natural to consider a *query-dependent* similarity measure which is specific to each query. Recent works [32, 9, 28, 24] have demonstrated the effectiveness of learning query-dependent similarity measures. In order to apply a query-dependent similarity measure in a large scale image retrieval system, however, we need to address the following two main issues.

(a) motorbike (b) firework

Figure 1. Different queries favor different similarity measures.

First, most of the efficient indexing schemes do not support query-dependent similarity measures. The commonly used indexing methods [26, 20, 16, 25, 30, 10, 31] organize all database images using a pre-defined, fixed similarity measure. Some works [9, 28] perform query adaptation on the top returned results. This second stage post-processing is, however, sub-optimal. It is still unclear how to build an efficient index with a query-dependent similarity measure.

Second, there is usually no or very limited training samples for learning a similarity measure for each online query. Existing works commonly choose the queries and their variants [28, 32, 24] as positive samples. However, generating negative samples requires either user interactions [28, 32] or time-consuming mining [24], which greatly affect the user experience.

To address the first issue, we introduce the use of a bilinear similarity model which expresses the similarity between two images in a bilinear form. This model allows the similarity be computed by first transforming one image by a linear transformation and then evaluating its Euclidian distance to the other image. To achieve search by query-dependent similarity measure, a query image is first transformed by a query-dependent transformation and the resulting *surrogate query* can then be used with any existing efficient indexing/retrieval methods.

*This work is done when Zhanghui Kuang is an intern at Microsoft Research Asia.

[1]We use "similarity" and "distance" interchangeably if there is no confusion. Note that, in this paper, we focus on similar image retrieval but not instance/object image retrieval.

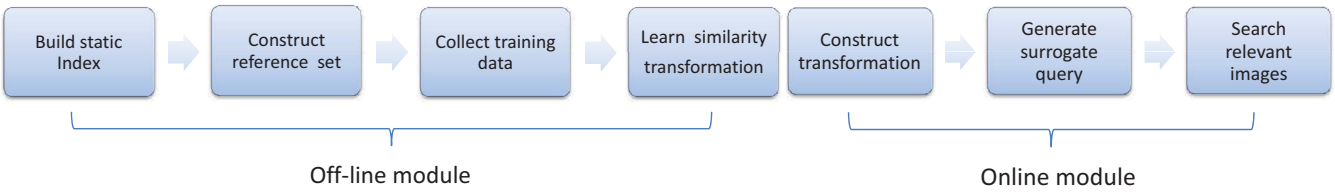


Figure 2. The work flows for two modules in our system.

To address the second issue, we leverage a reference set consisting of a number of reference queries. We collect training samples and learn a linear transformation for each reference query off-line. The transformation for an online query is then approximated by a linear combination of the transformations associated with its nearest neighbors in the reference set. To reduce the risk of over-fitting under limited training samples (especially negative samples), we propose a novel *angular regularization* to encourage the learned transformation to be not too far away from an identity matrix (i.e., we respect the original similarity to a certain extent). The learning is formulated as a quadratic optimization problem with only box bound constraints, and can be solved efficiently by a SMO-type algorithm.

The main contributions of this paper are:

1. We introduce the use of a bilinear similarity model in a large scale image retrieval system to achieve search by query-dependent similarity measure without sacrificing any efficiency of indexing and retrieval.

2. We propose an angular regularization for learning the bilinear similarity measure that can greatly reduce the risk of over-fitting under limited training data.

Although we focus on the large scale image retrieval problem, we believe our regularized, query-dependent bilinear similarity measure is quite general and can be applied to many other problems which need query adaptation.

2. Related Work

In the literature, there exist research studies that employ query-dependent similarity measures for ranking or re-ranking. Cui *et al*. [9] learned a similarity measure for each intention category to re-rank a short list of returned images. However, the category specific similarity measure is too complex to be used to rank images in the whole database. Arandjelović and Zisserman [2] learned a discriminative SVM with training samples generated by spatial geometry verification. In cases where holistic feature representations (*e.g*., color histogram) are used, geometry verification (and hence their method) is inapplicable. Shrivastava *et al*. [24] trained an examplar SVM for each query online. Their hard-negative mining, however, is too expensive (about three minutes on a 200-node cluster) for many practical applications. Relevance feedback [28, 32] can also be regarded as a form of learning query-dependent similarity measures. However, it requires user feedback to learn the ranking similarity measure online. On the contrary, our proposed approach does not need any user feedback online. Learning query-dependent similarity measures are also found in related fields, such as web page retrieval [12, 15, 3]. In contrast to our proposed approach, they index web pages by keywords without similarity of ranking, which makes the problem different.

3. System Overview

Our system consists of an off-line module to learn a bilinear similarity measure for each query in the reference set, and an online module to perform seach by a query-dependent similarity measure for each online query.

As shown in Fig. 2, the off-line module has four stages: 1) build a static index for images in the database using any indexing method (we use k-mean trees [21, 16] in this paper); 2) construct a reference query set (we run simple k-mean clustering on images in the database and use the cluster centers as the reference set); 3) collect training data for each reference query (each reference query is queried in the database, and the top returned images are manually labeled as either "relevant" or "not relevant"); 4) learn a bilinear similarity transformation for each reference query using the labeled training samples from the previous stage. Note that similarity between images in the off-line module are measured by Euclidean distance.

In the online module, we first construct a bilinear similarity transformation for an online query as a linear combination of the normalized transformations associated with its M-nearest neighbors in the reference set. We then transform the query into its surrogate query which is queried in the database.

4. Query-Dependent Similarities

As mentioned in Section 3, the core of our proposed approach is to learn one similarity measure for each reference query. In this section, we first study the problem of using the most common similarity measure, namely the *Mahalanobis distance*, in a search by query-dependent similarity measure scenario. We then introduce a regularized bilinear similarity measure which does not suffered from the mentioned

problem and can be learned efficiently by a SMO-type algorithm.

4.1. Mahalanobis Distance

In image retrieval, Mahalanobis distance [10, 16] is commonly used to measure similarity between images. Let an image be represented by a d-dimensional feature vector $\mathbf{x}_i$, and let $D = \{\mathbf{x}_i\}$ denote the image database. The Mahalanobis distance between a query image $\mathbf{x}_q$ and a database image $\mathbf{x}_i \in D$ is defined as $d_{\mathbf{M}}(\mathbf{x}_q, \mathbf{x}_i) = (\mathbf{x}_q - \mathbf{x}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_q - \mathbf{x}_i)$, where $\mathbf{M}$ is a positive definite matrix. When $\mathbf{M}$ is an identity matrix, Mahalanobis distance reduces to Euclidean distance. In a search by query-dependent similarity measure scenario, different queries have different similarity measures (*i.e.*, different $\mathbf{M}$), which require building different indexes for fast retrieval. However, this is not feasible in practice as the online queries and hence their similarity measures are not known beforehand, and it is not sensible or even possible to exhaust the space of similarity measures and build an index for each measure. If only one index is built, the returned images for ranking may contain few relevant images and this results in a low recall rate.

4.2. Bilinear Similarities

To avoid the aforementioned indexing problem, we consider a bilinear similarity model [7, 8]. The bilinear similarity between a query image $\mathbf{x}_q$ and a database image $\mathbf{x}_i \in D$ is defined as

$$\begin{aligned} s_{\mathbf{W}_q}(\mathbf{x}_q, \mathbf{x}_i) &= \mathbf{x}_q^{\mathrm{T}}\mathbf{W}_q\mathbf{x}_i \\ &\propto \hat{\mathbf{x}}_q^{\mathrm{T}}\mathbf{x}_i \\ &= \frac{1}{2}(\hat{\mathbf{x}}_q^{\mathrm{T}}\hat{\mathbf{x}}_q + \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \|\hat{\mathbf{x}}_q - \mathbf{x}_i\|^2) \\ &= 1 - \frac{1}{2}d_{\mathbf{I}}(\hat{\mathbf{x}}_q, \mathbf{x}_i), \end{aligned} \tag{1}$$

where $\hat{\mathbf{x}}_q = (\mathbf{W}_q^{\mathrm{T}}\mathbf{x}_q)/\|\mathbf{W}_q^{\mathrm{T}}\mathbf{x}_q\|$ and $d_{\mathbf{I}}(\hat{\mathbf{x}}_q, \mathbf{x}_i)$ is the Euclidean distance between $\hat{\mathbf{x}}_q$ and $\mathbf{x}_i$. Here, we assume $\mathbf{x}_i$ is L2-normalized [2]. Finding an image $\mathbf{x}_i$ that is most similar to the query $\mathbf{x}_q$ under the query-dependent measure $s_{\mathbf{W}_q}$ is therefore equivalent to finding an image $\mathbf{x}_i$ that is closest to $\hat{\mathbf{x}}_q$ in terms of Euclidean distance. It follows that we can build a static index for the database images using Euclidean distance and use it for fast retrieval without inducing any efficiency loss. The only modification we need before carrying out the actual query is to transform the query $\mathbf{x}_q$ into $\hat{\mathbf{x}}_q$ by the query-dependent similarity measurement matrix $\mathbf{W}_q$. We refer $\hat{\mathbf{x}}_q$ to as the *surrogate query* of $\mathbf{x}_q$. Fig. 3 illustrates how the surrogate query works. Since the query-dependent similarity measure is more specific to the query than the indexing similarity measure, the surrogate query has more relevant images than the original query in their corresponding neighborhoods. This has been validated in our experiments.

[2] This is a common practice and hence not a restriction.

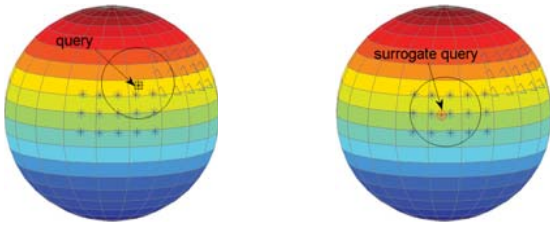


Figure 3. The symbols $*$ and $\triangle$ represent two types of feature points on an unit ball. ⊛ denotes the original query which is relevant to $*$, and its surrogate query is denoted by ⊛. The circles denote the neighborhoods of the original query and its surrogate query respectively.

4.3. Learning Bilinear Similarity Measure

The bilinear similarity measure can be learned by exploiting the relative similarities of image pairs generated from the training data collected in Step 3 of the off-line module. Formally, given a query $\mathbf{x}_q$, we form a set of triplets $T = \{(q, i, j)\}$ which depicts $\mathbf{x}_q$ and $\mathbf{x}_i$ are more similar than $\mathbf{x}_q$ and $\mathbf{x}_j$. Let the number of triplets be n (*i.e.*, $|T| = n$). We formulate the learning problem of the bilinear similarity measure as follows:

$$\begin{aligned} \min_{\mathbf{W}_q, \xi_{q,i,j}} \quad & \frac{1}{2}\|\mathbf{W}_q\|_{\mathrm{F}}^2 + C\sum_{q,i,j}\xi_{q,i,j} \\ s.t. \quad & \mathbf{x}_q^{\mathrm{T}}\mathbf{W}_q(\mathbf{x}_i - \mathbf{x}_j) \geqslant 1 - \xi_{q,i,j}, \\ & \xi_{q,i,j} \geqslant 0,\ \forall (q,i,j) \in T, \end{aligned} \tag{2}$$

where $\|\cdot\|_{\mathrm{F}}^2$ indicates squared Frobenius norm and $\xi_{q,i,j}$ are slack variables which add robustness for noisy training data. C is a trade-off parameter used to balance the margin regularization term $\|\mathbf{W}_q\|_{\mathrm{F}}^2$ and the hinge loss term $\sum \xi_{q,i,j}$. Fixing C to 1 gave good performance in our experiments.

Note that when $\mathbf{W}_q$ is a diagonal matrix, the bilinear similarity model becomes the Ranking SVM model [18]:

$$\begin{aligned} \min_{\mathbf{w}_q, \xi} \quad & \frac{1}{2}\mathbf{w}_q^{\mathrm{T}}\mathbf{w}_q + C\mathbf{I}_n^{\mathrm{T}}\xi \\ s.t. \quad & \mathbf{X}^{\mathrm{T}}\mathbf{w}_q \geqslant \mathbf{I}_n - \xi,\ \xi \geqslant \mathbf{0}, \end{aligned} \tag{3}$$

where $\mathbf{w}_q$ is a d-vector composed of the diagonal elements of $\mathbf{W}_q$, $\mathbf{I}_n$ is a n-vector with all elements being 1, ξ is a n-vector composed of $\xi_{q,i,j}$, and $\mathbf{X}$ is a $d \times n$ matrix with $(\mathbf{x}_i - \mathbf{x}_j) \circ \mathbf{x}_q$ as columns [3]. From the perspective of Ranking SVM, $\mathbf{x}_i \circ \mathbf{x}_q$ are referred to as features. They are with a special structure so that (3) is a bilinear model. For other features (*e.g.*, $(\mathbf{x}_q - \mathbf{x}_i) \circ (\mathbf{x}_q - \mathbf{x}_i)$), Ranking SVM does not

[3] $\mathbf{a} \circ \mathbf{b}$ represents the Hadamard product between $\mathbf{a}$ and $\mathbf{b}$.

belong to the bilinear similarity model. To summarize, not all Ranking SVMs belong to the bilinear similarity model, whereas the *diagonal* bilinear similarity model is a special kind of Ranking SVMs. In this paper, we adopt the diagonal bilinear similarity model for its simplicity.

4.4. Angular Regularization

Collecting sufficient training samples for each reference query is both laborious and time-consuming. Meanwhile, similarity measure learned from limited training samples suffers from the over-fitting problem and has poor generalization ability. In this paper, we introduce a novel angular regularization to tackle the issue of limited training samples.

In image retrieval, a similarity measure with $\mathbf{w}_q = \mathbf{I}_d$ (*i.e.*, cosine similarity) performs reasonably well in most cases. Therefore, similarity measures with $\mathbf{w}_q$ deviating slightly from $\mathbf{I}_d$ are desirable. A straightforward approach to reduce the risk of over-fitting under limited training samples is therefore to regularize (3) by $\|\mathbf{w}_q - \mathbf{I}_d\|$ [11]. Note that $\mathbf{w}_q$ and $s\mathbf{w}_q$, where s is an arbitrary positive scalar, are equivalent under the bilinear similarity model. Therefore, the angle between $\mathbf{w}_q$ and $\mathbf{I}_d$ (denoted by $\langle \mathbf{w}_q, \mathbf{I}_d \rangle$) is much more crucial than the magnitude of their difference. This angle can be measured by the minus cosine value:

$$-\frac{\mathbf{w}_q^{\mathrm{T}}\mathbf{I}_d}{\|\mathbf{w}_q\|\,\|\mathbf{I}_d\|}. \tag{4}$$

The smaller (4) is, the smaller the angle is and the more desirable $\mathbf{w}_q$ is. However, (4) is non-convex and regularizing (3) by (4) leads to local minima problem.

By Cauchy-Schwarz inequality [27], we have

$$-\|\mathbf{w}_q\|\|\mathbf{I}_d\| \leqslant \mathbf{w}_q^{\mathrm{T}}\mathbf{I}_d \leqslant \|\mathbf{w}_q\|\|\mathbf{I}_d\|. \tag{5}$$

We propose a regularizer that minimizes

$$(\|\mathbf{w}_q\|\,\|\mathbf{I}_d\|)^2 - (\mathbf{w}_q^{\mathrm{T}}\mathbf{I}_d)^2 = (\|\mathbf{w}_q\|\,\|\mathbf{I}_d\|)^2(1-\cos^2\langle \mathbf{w}_q, \mathbf{I}_d\rangle). \tag{6}$$

Minimizing (6) encourages a small $\|\mathbf{w}_q\|$ or a large $|\cos\langle \mathbf{w}_q, \mathbf{I}_d\rangle|$, which corresponds to $\langle \mathbf{w}_q, \mathbf{I}_d\rangle$ being close to 0 or π. Since a bilinear similarity measure with $\mathbf{w}_q = \mathbf{I}_d$ works well for most data, $\langle \mathbf{w}_q, \mathbf{I}_d\rangle$ is more likely to be close to 0. The learning problem in (3) can now be reformulated as:

$$\begin{aligned} \min_{\mathbf{w}_q,\xi}\ & \tfrac{1-\sigma}{2}\mathbf{w}_q^{\mathrm{T}}\mathbf{w}_q + \tfrac{\sigma}{2d}\left((\|\mathbf{w}_q\|\,\|\mathbf{I}_d\|)^2 - (\mathbf{w}_q^{\mathrm{T}}\mathbf{I}_d)^2\right) + C\mathbf{I}_n^{\mathrm{T}}\xi \\ s.t.\ & \mathbf{X}^{\mathrm{T}}\mathbf{w}_q \geqslant \mathbf{I}_n - \xi,\ \ \xi \geqslant \mathbf{0}, \end{aligned} \tag{7}$$

where $\sigma \in [0, 1)$ is a trade-off parameter used to balance the impact of the margin regularizer and that of the angular regularizer. Note that the impact of the angular regularizer decreases with σ, and (7) reduces to (3) when $\sigma = 0$.

Fig. 4 demonstrates how σ impacts the learning results. We learned a query-dependent similarity measure using the same set of training samples (we labeled 10-nearest neighbors of one query under Euclidean distance) but under different values of σ. It can be observed that $\langle \mathbf{w}_q, \mathbf{I}_d\rangle$ becomes smaller and smaller as σ increases, whereas the magnitude of $\mathbf{w}_q$ becomes larger and larger. We carried out the query with the learned similarity measures and the top-10 ranked images are shown in Fig. 4. It can be seen that the similarity measures learned with the angular regularization performed much better than that without it. This example shows that, with the angular regularization, the bilinear similarity measure can be learned effectively from very few training samples.

4.5. Optimization Methods

Although the regularized bilinear model in (7) is convex, its energy function is not differentiable at some points as it contains the hinge loss term. In order to optimize it efficiently, we consider its dual problem. Let $\mathbf{A} = \mathrm{diag}(\mathbf{I}_d) - \frac{\sigma}{d}\mathbf{I}_d\mathbf{I}_d^{\mathrm{T}}$, where $\mathrm{diag}(\mathbf{I}_d)$ denotes a square matrix with the elements of $\mathbf{I}_d$ on its diagonal. Eq. (7) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{w}_q,\xi}\ & \frac{1}{2}\mathbf{w}_q^{\mathrm{T}}\mathbf{A}\mathbf{w}_q + C\mathbf{I}_n^{\mathrm{T}}\xi, \\ s.t.\ & \mathbf{X}^{\mathrm{T}}\mathbf{w}_q \geqslant \mathbf{I}_n - \xi,\ \xi \geqslant \mathbf{0}. \end{aligned} \tag{8}$$

Note that $\mathbf{w}_q^{\mathrm{T}}\mathbf{A}\mathbf{w}_q \geqslant 0$ for all $\sigma \in [0, 1)$. The equality holds if and only if $\mathbf{w}_q = \mathbf{0}$. Therefore, $\mathbf{A}$ is a positive-definite matrix and its inverse $\mathbf{A}^{-1}$ exists. Introducing the non-negative Lagrange multipliers α and μ for the inequalities $\mathbf{X}^{\mathrm{T}}\mathbf{w}_q \geqslant \mathbf{I}_n - \xi$ and $\xi \geqslant 0$, respectively, gives

$$L = \frac{1}{2}\mathbf{w}_q^{\mathrm{T}}\mathbf{A}\mathbf{w}_q + C\mathbf{I}_n^{\mathrm{T}}\xi + \alpha^{\mathrm{T}}(\mathbf{I}_n - \xi - \mathbf{X}^{\mathrm{T}}\mathbf{w}_q) - \mu^{\mathrm{T}}\xi. \tag{9}$$

Taking the derivatives of L with respective to $\mathbf{w}_q$ and ξ, and setting them to zero gives $\mathbf{w}_q = \mathbf{A}^{-1}\mathbf{X}\alpha$ and the dual problem:

$$\min\ \frac{1}{2}\alpha^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{X}\alpha - \mathbf{I}_n^{\mathrm{T}}\alpha,\ s.t.\ \mathbf{0} \leqslant \alpha \leqslant C\mathbf{I}_n. \tag{10}$$

The above dual problem is a Quadratic Programming (QP) problem with only box bound constraints, and can be solved by off-the-shelf solvers [29, 13]. When the number of variables is large, decomposition algorithms [17, 6, 22] are preferred. Chang and Lin [6] utilized SMO [22] to solve a quadratic problem generated by SVM by selecting two variables in each iteration. However, the algorithm which is designed for QPs with equality constraints cannot be directly used to solve (10). Inspired by these work, a SMO-type approach is proposed. We optimize one variable instead of two in each iteration. Each subproblem is a simple QP with one variable and hence can be solved efficiently with an analytical solution.

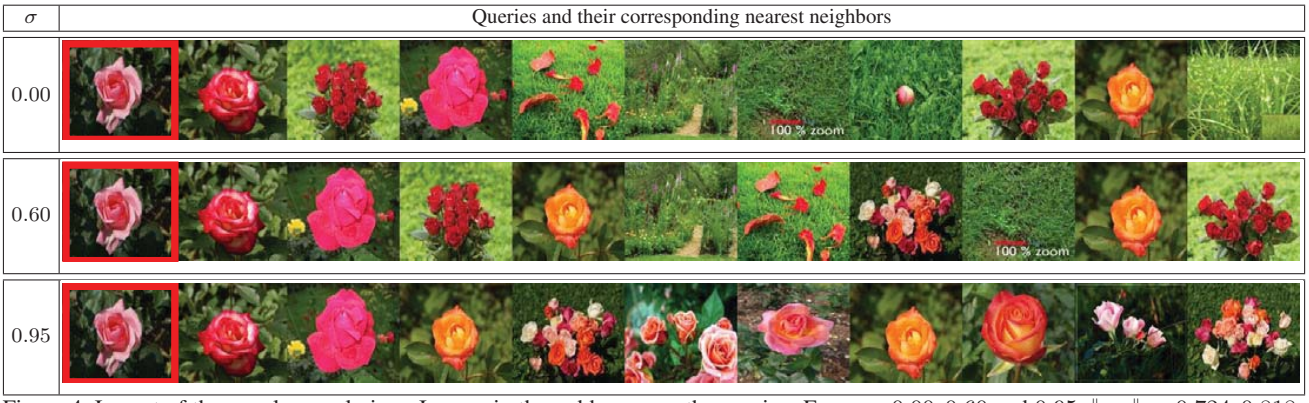


Figure 4. Impact of the angular regularizer. Images in the red boxes are the queries. For $\sigma = 0.00, 0.60$ and 0.95, $\|\mathbf{w}_q\| = 0.724, 0.818$ and 3.397, and $\langle \mathbf{w}_q, \mathbf{I}_d \rangle = 1.339, 0.959$ and 0.209 radians respectively.

Let $\mathbf{Q} = \mathbf{X}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{X}$ and $\nabla f(\alpha)$ be the gradient of (10). According to the Karush-Kuhn-Turcker (KKT) conditions [5] of (10), the stopping criteria are given by:

$$\nabla_i f(\alpha) \begin{cases} \leqslant 0 & \text{if } \alpha_i = C \\ = 0 & \text{if } 0 < \alpha_i < C \\ \geqslant 0 & \text{if } \alpha_i = 0. \end{cases} \tag{11}$$

In each iteration, α_i with the largest absolute gradient $|\nabla_i f(\alpha)|$ in the variable subset, whose members violate the stopping criteria in (11), is selected as the active variable. We then solve the following subproblem:

$$\begin{aligned} &\min_{\alpha_i} \ \tfrac{1}{2}\mathbf{Q}_{ii}\alpha_i^2 + (\mathbf{Q}_{iB}\alpha_B - 1)\alpha_i + \tfrac{1}{2}\alpha_B^{\mathrm{T}}\mathbf{Q}_{BB}\alpha_B - \mathbf{I}_{n-1}^{\mathrm{T}}\alpha_B \\ &s.t. \quad 0 \leqslant \alpha_i \leqslant C, \end{aligned} \tag{12}$$

where $B = \{1, 2, \ldots, n\} \backslash \{i\}$. α_B denotes the set of non-active variables. This subproblem has an analytical solution given by

$$\alpha_i = \begin{cases} 0 & \text{if } m < 0 \\ m & \text{if } 0 \leqslant m \leqslant C \\ C & \text{if } m > C \end{cases} \tag{13}$$

where $m = \frac{1 - \mathbf{Q}_{iB}\alpha_B}{\mathbf{Q}_{ii} + \epsilon}$ and ϵ is a positive infinitesimal. The optimization procedure is summarized in Algorithm 1.

Algorithm 1 SMO-type Algorithm for solving (10)

1: Input $\mathbf{Q}$ and C. Set $t = 1$. Initialize $\alpha^t = \mathbf{0}$.
2: **repeat**
3: Find the active variable index i. Define non-active variable index set B.
4: Solve subproblem (12) by (13).
5: Set α_i^{t+1} to the optimal solution of (12), and $\alpha_B^{t+1} = \alpha_B^t$. $t = t + 1$.
6: **until** the stopping condition (11) holds.

5. Experimental Results

We evaluated our proposed method on three datasets: the MNIST dataset[4], the CIFAR-10 dataset [19], and our own dataset with 1 million images downloaded from the web. For the first two, we exhaustively searched query-relevant images to evaluate the performance of learning regularized query-dependent bilinear similarity measures. For the third one, the performance was evaluated in a large scale image retrieval scenario, and multi-probe k-mean trees were employed to search approximate nearest neighbors efficiently.

5.1. Evaluation Protocols

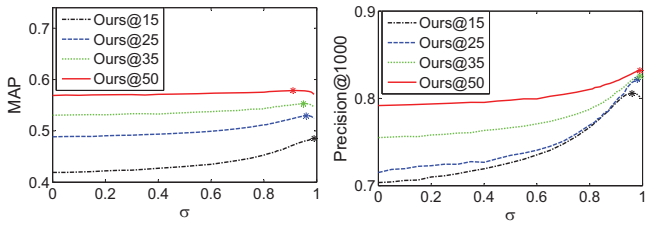
Image retrieval was evaluated by two commonly employed protocols, namely the Mean Average Precision (MAP) [1], (*i.e.*, the area under the recall precision curve) and the average precision of top-R ranked images for each test query [14] (denoted by Precision@R here).

We compared our regularized query-dependent bilinear similarity measure with the following two baseline methods and one state-of-the-art technique:

1. Euclidean: Euclidean distance is used to evaluate the similarities between query and database images.
2. Query-Independent Ranking SVM (QI-RSVM): Ranking SVM is used to learn a ranking function from the training samples of all reference queries, which is then applied to all queries.
3. Query-Dependent Ranking SVM (QD-RSVM): query-dependent Ranking SVMs are learned [12]. It is equivalent to our similarity measure with $\sigma = 0$. For a fair comparison, all other parameters were set the same for both QD-RSVM and our method.

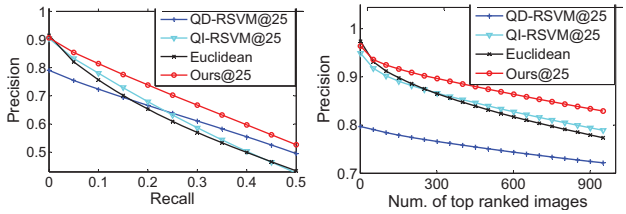
To show how the number of training samples for each reference query (N) affects our proposed method, we tested our method under different N settings (denoted by

[4] http://yann.lecun.com/exdb/mnist/



(a) MAP change (b) Precision@1, 000 change

Figure 5. Results on the MNIST dataset. (a) MAP under different values of N and σ. (b) Precision@1, 000 under different values of N and σ. "*" on each curve indicates the peak performance.



(a) Recall precision curve (b) Precision@R

Figure 6. Comparison with Euclidean, QI-RSVM and QD-RSVM on the MNIST dataset.

Ours@N). For comparison, we also tested QI-RSVM and QD-RSVM under different values of N (denoted by QI-RSVM@N and QD-RSVM@N respectively). 10-nearest reference queries are found to construct transformation for each online query in our experiments (*i.e.*, $M = 10$).

5.2. Results on the MNIST Dataset

The MNIST dataset contains 70k 28×28 aligned digital images. Each sample is associated with a label from 0 to 9. We rearranged the pixel intensities of each image into a 784-vector. These vectors were projected into a 260-dimensional subspace by PCA. 1,000 testing data were sampled. The rest of the data were clustered into 1,000 clusters, and the medoids were selected as the reference queries.

Our proposed method was evaluated under different values of N and σ, and the results are shown in Fig. 5. Our method achieved the best MAP at $\sigma = 0.910, 0.950, 0.960, 0.992$, and the best precision@1, 000 at $\sigma = 0.993, 0.991, 0.970, 0.960$, respectively, for $N = 50, 35, 25, 15$. As more and more training samples were available (*i.e.*, as N increased), the gain arose from the angular regularization became smaller and smaller. In a large scale image retrieval scenario, collection of sufficient training samples is both laborious and time-consuming, and sometimes even impossible. We believe that the proposed angular regularization is significantly useful since it enables the similarity measures to be learned effectively from limited training data. As our proposed method works well with $\sigma \in (0.9, 1)$ for different values of N, we will report its performances at $\sigma = 0.95$ for the rest of this paper.

Table 1. Ranking performance on the MNIST dataset measured by MAP.

Methods \ N	15	25	35	50
Ours@N	**0.480**	**0.529**	**0.558**	**0.578**
QD-RSVM@N	0.419	0.488	0.530	0.570
QI-RSVM@N	0.450	0.454	0.468	0.476
Euclidean	0.462			

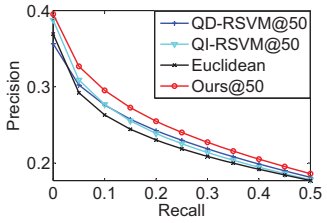


Figure 7. Performance on the CIFAR-10 dataset.

Fig. 6 compares the results of our proposed method with that of Euclidean, QI-RSVM, and QD-RSVM in terms of recall precision curves and precision@R. Our proposed method clearly outperformed the others. Specifically, the gains in precision@300 are about 3.64%, 3.51%, and 17.05% over Euclidean, QI-RSVM, and QD-RSVM respectively.

As shown in Table 1, our proposed method achieved the highest search accuracy in terms of MAPs. The gain of Ours@50 is around 21% to 25% over Euclidean and QI-RSVM@50. The MAP of our method is 14.56%, 8.40%, 5.28% and 1.40% higher than that of its competitor QD-RSVM when $N = 15, 25, 35$ and 50 respectively.

5.3. Results on the CIFAR-10

The CIFAR-10 dataset contains 60k 32×32 colour images in 10 classes, with 6,000 images per class. For each image, we extracted a 320-dimensional Gist descriptor computed at 3 scales with 8, 8, and 4 orientations respectively. These 320-vectors were projected into a 225-dimensional subspace by PCA. 10k images were sampled as testing data, while the rest were clustered into $2, 250$ clusters and with their medoids selected as the reference queries.

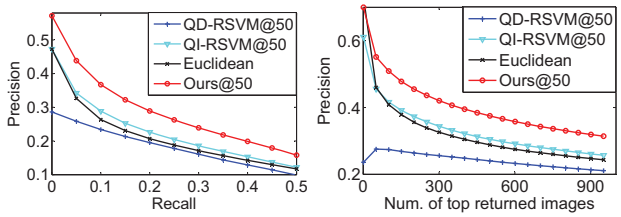
Fig. 7 compares the performance of our proposed method on the CIFAR-10 dataset with that of Euclidean, QI-RSVM, and QD-RSVM in terms of recall precision curves. The MAPs for our proposed method, Euclidean, QI-RSVM, and QD-RSVM are 0.205, 0.189, 0.193, and 0.194 respectively.

5.4. Results on the Web Image Dataset

About 70k images were crawled from the Google image search engine using 17 keywords. Non-relevant images were filtered out manually. Some random images were also crawled as background images to make up the total number

Table 2. Recalls on the web image dataset under different probe number K settings.

Methods \ K	5	10	15	20
Ours@50	**0.371**	**0.451**	**0.506**	**0.545**
QD-RSVM@50	0.221	0.275	0.321	0.350
QI-RSVM@50	0.311	0.409	0.474	0.518
Euclidean	0.328	0.412	0.468	0.511



(a) Recall precision curve (b) Precision@R

Figure 8. Performance on the web image dataset.

of images to 1 million.

Color histogram, Gist, PHoG [4] and spatial color histogram [23] were extracted from each image to form a feature vector, which was then projected into a 60-dimensional subspace by PCA. 360 non-background images were sampled as testing queries, and the reference query set was constructed from $2,000$ medoids of the non-background images. We evaluated our proposed method on this web image dataset, and compared the results with that of Euclidean, QI-RSVM, and QD-RSVM.

We clustered the dataset into 1,000 clusters with different initializations and then built three k-mean trees. The number of probes (*i.e.*, the number of buckets from one k-mean tree returned for each query) was $K \in [5, 10, 15, 20]$. The method based on Euclidean distance used the original queries, while others used the surrogate queries. Table. 2 compares the proposed method with others in terms of recall rates under different values of K. QD-RSVM@50 performed the worst because its similarity measures learned had poor generalization ability, and hence the surrogate queries were not good. Ours@50 achieved recall rates higher than those of Euclidean by 13.11%, 9.47%, 8.12%, and 6.65% when $K = 5, 10, 15$ and 20 respectively. This indicates that simple applying one static index without query-dependent transformation does sacrifice the retrieval recall.

Fig. 8 (a) shows the recall precision curves. It can be seen that our proposed method is far superior to others. The MAPs for Ours@50, Euclidean, QI-RSVM@50, and QD-RSVM@50 are 0.189, 0.139, 0.146 and 0.117 respectively. The very interesting observation is that the performance of QD-RSVM@50 is even worse than that of Euclidean. This may be due to insufficient training data. Fig. 8 (b) shows the top ranked images precision. Our method greatly outperformed its competitors. For example, the precision of the Ours@50 for $R = 50$ is 0.551 while that of Euclidean, QI-RSVM@50, and QD-RSVM@50 are 0.460, 0.456, and 0.276 respectively. One interesting phenomenon is that the precision of QD-RSVM@50 for $R = 1$ is lower than that for $R = 50$. This can be explained as follows: the similarity model employed by QD-RSVM is the bilinear model which cannot guarantee that $s_W(\mathbf{x}_q, \mathbf{x}_q) \geqslant s_W(\mathbf{x}_q, \mathbf{x}_i)$ for all i. Therefore, its precision@1 may be extremely low. Although the proposed method also use the bilinear similarity model, the angular regularization forces the model to be close to Euclidean distance. Since $-d_{\mathbf{I}}(\mathbf{x}_q, \mathbf{x}_q) \geqslant -d_{\mathbf{I}}(\mathbf{x}_q, \mathbf{x}_i)$ for all i, our proposed method can avoid this problem.

To visualize the quality of the nearest neighbors found, we show top-25 neighbors retrieved by different techniques for four example queries in Fig. 9.

6. Conclusion

In this paper, we introduce a regularized query-dependent bilinear similarity measure for large scale image retrieval. The proposed bilinear model allows the search to be carried out using a surrogate query with a static index. This makes search by query-dependent similarity measure possible without sacrificing any efficiency of indexing and retrieval. To tackle the problem of limited training samples, the similarity transformation of an online query is approximated by a linear combination of the similarity measure matrices associated with its nearest neighbors in a reference query set. A novel angular regularization constraint is proposed to avoid the over-fitting problem in learning the similarity measure for each reference query with limited training data. Experimental results on two public datasets and our 1-million web-image dataset demonstrate that our proposed method outperforms other state-of-the-art methods in terms of MAPs, precisions and recall rates.

References

[1] R. Arandjelović and A. Zisserman. Smooth object retrieval using a bag of boundaries. In *ICCV*, 2011.

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012.

[3] J. Bian, T.-Y. Liu, T. Qin, and H. Zha. Ranking with query-dependent loss for web search. *WSDM*, pages 141–150, 2010.

[4] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*, 2007.

[5] L. Boyd, Stephen Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

[6] C.-c. Chang and C.-j. Lin. LIBSVM : a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):1–27, 2011.

[7] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.*, 11:1109–1135, 2010.

[8] K. Crammer and G. Chechik. Adaptive regularization for weight matrices. In *ICML*, 2012.

[9] J. Cui, F. Wen, and X. Tang. Real time Google and live image search re-ranking. In *ACM MM*, pages 133–142, 2008.

[10] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *SoSCG*, pages 253–262, 2004.

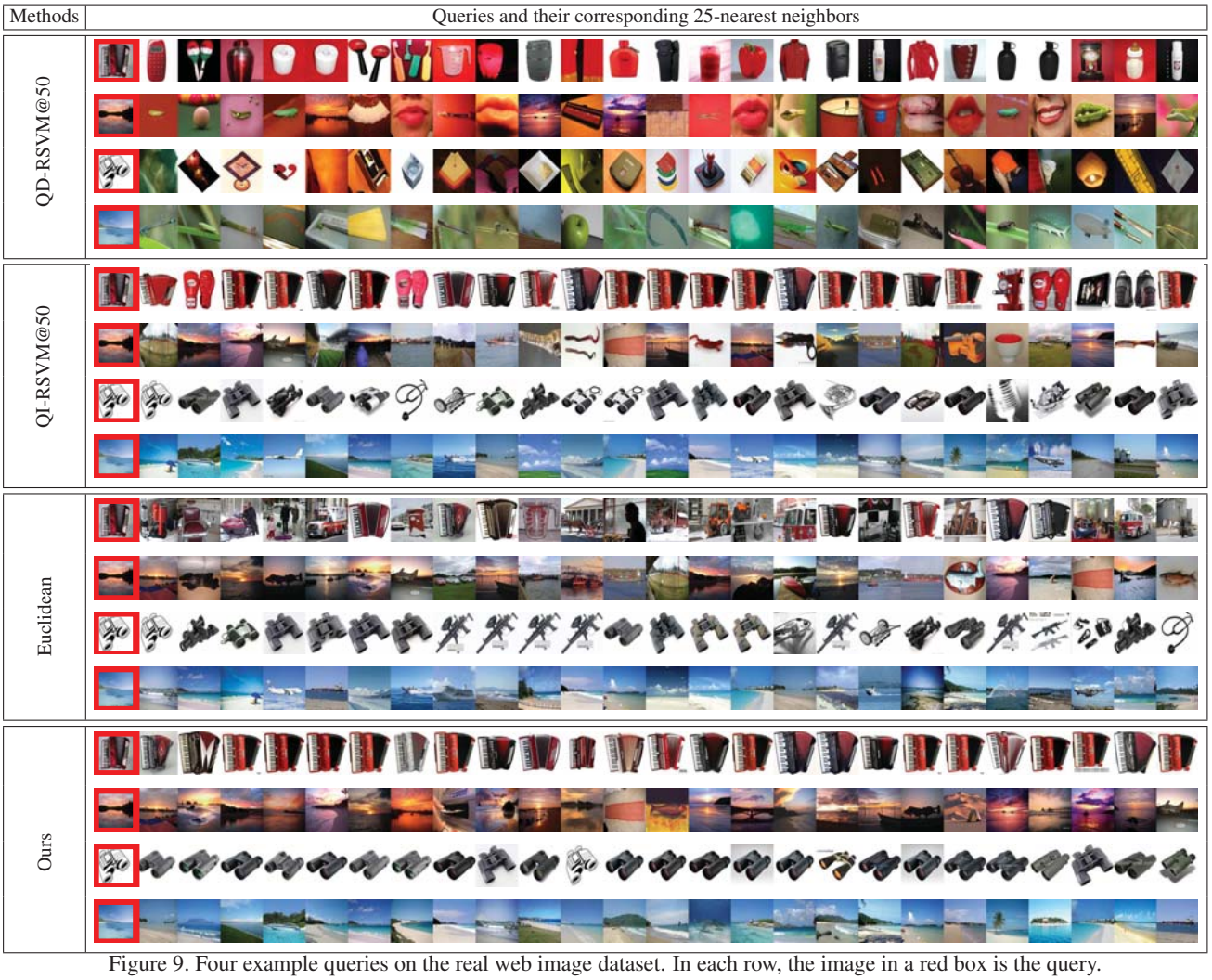


Figure 9. Four example queries on the real web image dataset. In each row, the image in a red box is the query.

[11] B. Geng, L. Yang, C. Xu, and X.-S. Hua. Ranking model adaptation for domain-specific search. *ACM CIKM*, pages 197–206, 2009.

[12] X. Geng, T. Liu, T. Qin, A. Arnold, H. Li, and H.-Y. Shum. Query dependent ranking using k-nearest neighbor. In *SIGIR*, pages 115–122, 2008.

[13] E. M. Gertz and S. J. Right. Object-oriented software for quadratic programming. *ACM Transactions on Mathematical Software*, 29:58–81, 2003.

[14] Y. Gong and S. Lazebnik. Iterative quantization : a procrustean approach to learning binary codes. In *CVPR*, pages 817–824, 2011.

[15] V. Jain and M. Varma. Learning to re-rank : query-dependent image re-ranking using click data. In *WWW*, pages 277–286, 2011.

[16] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, 33(1):117–128, Jan. 2011.

[17] T. Joachims. Making large-scale support vector machine learning practical. In *Advances in kernel methods*, pages 169–184. 1999.

[18] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, pages 133–142, 2002.

[19] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[20] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[21] L. Paulevé, H. Jégou, and L. Amsaleg. Locality sensitive hashing: a comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31(11):1348–1358, Aug. 2010.

[22] J. C. Platt. Sequential minimal optimization: a fast algorithm for training support vector machines. Technical report, Microsoft Research, 1998.

[23] A. R. Rohini, R. K. Srihari, and Z. Zhang. Spatial color histograms for content-based image retrieval. In *International Conference on Tools with Artificial Intelligence*, pages 183–186, 1999.

[24] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. In *SIGGRAPH Asia*, 2011.

[25] C. Silpa-anan and R. Hartley. Optimised KD-trees for fast image descriptor matching. In *CVPR*, 2008.

[26] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003.

[27] J. M. Steele. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities.* Cambridge University Press, 2004.

[28] D. Tao, X. Li, S. J. Maybank, and S. Member. Negative samples analysis in relevance feedback. *IEEE Transactions on Knowledge and Data Engineering*, 19(4):568–580, 2007.

[29] R. J. Vanderbei. LOQO: an interior point code for quadratic programming. *Optimization Methods and Software*, 11(1):451–484, 1999.

[30] N. Verma, S. Kpotufe, and S. Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? In *UAI*, 2009.

[31] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, number 1, pages 1–8, 2008.

[32] J. Zhang and L. Ye. Local aggregation function learning based on support vector machines. *Signal processing*, 89(11):2291–2295, Nov. 2009.