Output Regularized Metric Learning with Side Information

Wei Liu[1], Steven C.H. Hoi[2], and Jianzhuang Liu[1]

[1] Department of Information Engineering, The Chinese University of Hong Kong
Hong Kong, China
{wliu5,jzliu}@ie.cuhk.edu.hk
[2] School of Computer Engineering,
Nanyang Technological University, Singapore
chhoi@ntu.edu.sg

Abstract. Distance metric learning has been widely investigated in machine learning and information retrieval. In this paper, we study a particular content-based image retrieval application of learning distance metrics from historical relevance feedback log data, which leads to a novel scenario called *collaborative image retrieval*. The log data provide the *side information* expressed as relevance judgements between image pairs. Exploiting the side information as well as inherent neighborhood structures among examples, we design a convex regularizer upon which a novel distance metric learning approach, named *output regularized metric learning*, is presented to tackle collaborative image retrieval. Different from previous distance metric methods, the proposed technique integrates synergistic information from both log data and unlabeled data through a regularization framework and pilots the desired metric toward the ideal output that satisfies pairwise constraints revealed by side information. The experiments on image retrieval tasks have been performed to validate the feasibility of the proposed distance metric technique.

Keywords: Distance Metric Learning, Side Information, Output Regularized Metric Learning, Collaborative Image Retrieval.

1 Introduction

Recently, there are some emerging research interests in exploring the historical log data of the user's relevance feedback in content-based image retrieval (CBIR). Hoi *et al.* [1] proposed the log-based relevance feedback with support vector machines (SVMs) through engaging the feedback log data in traditional online relevance feedback sessions. In this paper, we study distance metric learning to discover the potential of the log data so that the needs of online relevance feedback can be avoided.

Distance metric learning has attracted increasing attention in recent machine learning and computer vision studies, which may be classified into two main categories. The first category is supervised learning approaches for classification, where distance metrics are usually learned from the training data associated

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part III, LNCS 5304, pp. 358–371, 2008.
© Springer-Verlag Berlin Heidelberg 2008

with explicit class labels. The representative techniques include Linear Discriminant Analysis (LDA) [2] and some other recently proposed methods, such as Neighbourhood Components Analysis (NCA) [3], Maximally Collapsing Metric Learning (MCML) [4], metric learning for Large Margin Nearest Neighbor classification (LMNN) [5], and Local Distance Metric Learning (LDML) [6].

Our work is closer to the second category, i.e., semi-supervised distance metric learning which learns distance metrics from pairwise constraints, or known as *side information* [7]. Each constraint indicates whether two data objects are "similar" (*must-link*) or "dissimilar" (*cannot-link*) in a particular learning task. A well-known metric learning method with these constraints was proposed by Xing *et al.* [7], who cast the learning task into a convex optimization problem and applied the generated solution to data clustering. Following their work, there are several emerging metric techniques in this "semi-supervised" direction. For instance, Relevance Component Analysis (RCA) learns a global linear transformation by exploiting only the equivalent (must-link) constraints [8]. Discriminant Component Analysis (DCA) improves RCA via incorporating the inequivalent (cannot-link) constraints [9]. Si *et al.* [10] proposed a regularized metric learning method by formulating the side information into a semidefinite program.

Particularly, we are aware that routine metric techniques may be sensitive to noise and fail to learn reliable metrics when handling small amount of side information. In this paper, we present a new semi-supervised distance metric learning algorithm to incorporate the unlabeled data together with side information in producing metrics with high fidelity. Specifically, we develop an output regularized framework to integrate the synergistic information from both the log data and the unlabeled data for the goal of coherently learning a distance metric. The proposed *output regularized metric learning* (ORML) algorithm is elegantly formulated, resulting in a close-form solution which can be obtained with a global optimum substantially efficiently.

2 Collaborative Image Retrieval

In the field of CBIR, choosing appropriate distance metrics plays a key role in establishing an effective CBIR system. Regular CBIR systems usually adopt Euclidean metrics for distance measure on images represented into vector form. Unfortunately, the Euclidean distance is generally not effective enough in retrieving relevant images. A main reason stems from the well-known semantic gap between low-level visual features and high-level semantic concepts [11].

To remedy the semantic gap issue, relevance feedback is frequently engaged in CBIR systems. Relevance feedback mechanism has been vastly studied in the CBIR community and demonstrated to improve retrieval performance. However, the relevance feedback mechanism has some drawbacks in practice. One problem is that relevance feedback often has to involve overloaded communication between systems and users, which might not be efficient for real-time applications. Further, relevance feedback often has to be repeated several times for retrieving relevant images. This procedure could be a tedious task for users.

Thus, relevance feedback may not be an efficient and permanent solution for addressing the semantic gap from a long-term perspective.

In this paper, we consider an alternative solution, called *collaborative image retrieval* (CIR), for attacking the semantic gap challenge by leveraging the historical log data during the user's relevance feedback. CIR has attracted a surge of research interests in the past few years [1,10]. The key to CIR is to find a convenient and effective way of leveraging the log data in relevance feedback so that the semantic gap can be successfully reduced. A lot of ways could be studied to use the log data to boost the retrieval performance. In this paper, we explore to learn distance metrics from the log data for image retrieval tasks, and address some practical problems in applying distance metric techniques to the CIR application.

3 Distance Metric Learning with Side Information

3.1 Side Information

Assume that we have a set of n data points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^m$, and two sets of pairwise constraints on these data points:

$$\begin{aligned} \mathcal{S} &= \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are judged to be equivalent}\} \\ \mathcal{D} &= \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are judged to be inequivalent}\}, \end{aligned} \tag{1}$$

where $\mathcal{S}$ is the set of similar pairwise constraints, and $\mathcal{D}$ is the set of dissimilar pairwise constraints. Each pairwise constraint indicates if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are equivalent (similar) or inequivalent (dissimilar) judged by users under certain application context. The two types of constraints $\mathcal{S}$ and $\mathcal{D}$ are referred to as side information. Note that it is not necessary for all the points in $\mathcal{X}$ involved in $\mathcal{S}$ or $\mathcal{D}$.

For any pair of points $\mathbf{x}_i$ and $\mathbf{x}_j$, let $d(\mathbf{x}_i, \mathbf{x}_j)$ denote the distance function between them. By introducing a symmetric matrix $A \in \mathbb{R}^{m \times m}$, we can then express the distance function as follows:

$$d_A(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_A = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T A (\mathbf{x}_i - \mathbf{x}_j)}. \tag{2}$$

In practice, the metric matrix A is a valid metric if and only if it satisfies the non-negativity and the triangle inequality conditions. In other words, A must be positive semidefinite, i.e., $A \succeq 0$. Generally speaking, the matrix A parameterizes a family of Mahalanobis distances defined on the vector space $\mathbb{R}^m$. As an extreme case, when setting A to be the identity matrix $I_{m \times m}$, the distance in Eqn. (2) becomes the common Euclidean distance.

Abiding by the settings of semi-supervised learning, our learning problem is to learn an optimal square matrix $A \in \mathbb{R}^{m \times m}$ from a collection of data points $\mathcal{X} \subset \mathbb{R}^m$ coupled with a set of similar pairwise constraints $\mathcal{S}$ and a set of dissimilar pairwise constraints $\mathcal{D}$. So far, the central theme to attack metric learning is to design an appropriate optimization objective and then find an efficient algorithm to solve the optimization problem.

3.2 Optimization Model

One intuitive yet effective principle for designing metric learning approaches is to minimize the distances between the data points with similar constraints and meanwhile to maximize the distances between the data points with dissimilar constraints. We call it as the *min-max* principle. Some existing work [7][10] can be interpreted in terms of the min-max principle.

To make metric learning techniques practical, the second principle we want to highlight is the *regularization* principle, which is a key to empowering the learnt metric with the generalization and robustness capabilities. Motivated by the idea of regularization in kernel machines [12], we formulate a general regularization prototype for distance metric learning as follows:

$$\begin{aligned} \min_{A} \quad & \mathcal{R}(A, \mathcal{X}, \mathcal{S}, \mathcal{D}) + \gamma \mathcal{V}(A, \mathcal{S}, \mathcal{D}) \\ s.t. \quad & A \succeq 0 \end{aligned} \tag{3}$$

where $\mathcal{R}(\cdot)$ is some regularizer defined on the target metric A, raw samples $\mathcal{X}$ and side information $\mathcal{S}$ and $\mathcal{D}$. $\mathcal{V}(\cdot)$ is some loss function defined on A and side information, and γ is a regularization parameter for controlling the trade-off between two terms in Eqn. (3). According to the min-max principle, a good loss function $\mathcal{V}(\cdot)$ should be designed in a way such that its minimization will simultaneously result in shrinking the distances between points with similar constraints and elongating the distances between points with dissimilar constraints.

3.3 Dissimilarity-Enhanced Regularizer

There are a lot of options to decide a regularizer in the above regularization prototype. The simplest one is based on the Frobenius norm: $\mathcal{R}(A) = \|A\|_{\mathrm{F}}$ that simply prevents any elements within the matrix A from being overlarge [10]. However, this regularizer cannot take advantage of any side information. Hence, we intend to formulate a better regularizer by exploiting side information and unlabeled data information which is beneficial to semi-supervised learning tasks.

Given the collection of n data points $\mathcal{X}$ including the unlabeled data and the side information $\mathcal{S}$ and $\mathcal{D}$, we define a weight matrix $W \in \mathbb{R}^{n \times n}$ on $\mathcal{X}$:

$$W_{ij} = \begin{cases} \alpha, & (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \\ \beta, & (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \\ 1, & (\mathbf{x}_i, \mathbf{x}_j) \notin \mathcal{S} \cup \mathcal{D} \text{ and } (\mathbf{x}_i \in \mathcal{N}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\mathcal{N}(\mathbf{x}_i)$ denotes the list composed of k nearest neighbors of the point $\mathbf{x}_i$, and α, $\beta > 0$ are two weighting parameters corresponding to $\mathcal{S}$, $\mathcal{D}$. It is worth mentioning that W absorbs and encodes the side information as well as inherent neighborhood structures among examples. We define another weight matrix $T \in \mathbb{R}^{n \times n}$ based on only dissimilarity constraints:

$$T_{ij} = \begin{cases} \beta, & (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

To delve into a metric matrix A, one can assume there exits a linear mapping $U : \mathbb{R}^m \mapsto \mathbb{R}^r$ to constitute $A = UU^T$, where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathbb{R}^{m \times r}$. We require $\mathbf{u}_1, \ldots, \mathbf{u}_r$ be linearly independent so that r is the rank of A. Then, the distance under A between two inputs can be written as:

$$d_A(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T A(\mathbf{x}_i - \mathbf{x}_j)} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T UU^T(\mathbf{x}_i - \mathbf{x}_j)}$$
$$= \left\| U^T(\mathbf{x}_i - \mathbf{x}_j) \right\| = \sqrt{\sum_{d=1}^{r} \left(\mathbf{u}_d^T \mathbf{x}_i - \mathbf{u}_d^T \mathbf{x}_j \right)^2}. \tag{6}$$

Minimizing $d_A(\mathbf{x}_i, \mathbf{x}_j)$ will lead to $\mathbf{u}_d^T \mathbf{x}_i - \mathbf{u}_d^T \mathbf{x}_j \to 0$ corresponding to projection direction $\mathbf{u}_d$. Specially, we define a new function as follows

$$h_A(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i + \mathbf{x}_j)^T A(\mathbf{x}_i + \mathbf{x}_j)} = \left\| U^T(\mathbf{x}_i + \mathbf{x}_j) \right\|$$
$$= \sqrt{\sum_{d=1}^{r} \left(\mathbf{u}_d^T \mathbf{x}_i + \mathbf{u}_d^T \mathbf{x}_j \right)^2}. \tag{7}$$

At this time, minimizing $h_A(\mathbf{x}_i, \mathbf{x}_j)$ will lead to $\mathbf{u}_d^T \mathbf{x}_i + \mathbf{u}_d^T \mathbf{x}_j \to 0$, which actually pushes $\mathbf{x}_i$ and $\mathbf{x}_j$ far away along each projection direction $\mathbf{u}_d$.

Intuitively, we would like to minimize $d_A(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ meet the similarity constraint or belong to the nearest neighbors of each other, and meanwhile to minimize $h_A(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ meet the dissimilarity constraint. By leveraging side information and neighborhood structures in the weight matrix W, we formulate the regularizer as follows:

$$\mathcal{R}(A, \mathcal{X}, \mathcal{S}, \mathcal{D}) = \frac{1}{2} \left[\sum_{(\mathbf{x}_i, \mathbf{x}_j) \notin \mathcal{D}} d_A^2(\mathbf{x}_i, \mathbf{x}_j) W_{ij} + \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} h_A^2(\mathbf{x}_i, \mathbf{x}_j) W_{ij} \right]$$
$$= \frac{1}{2} \sum_{d=1}^{r} \left[\sum_{(\mathbf{x}_i, \mathbf{x}_j) \notin \mathcal{D}} \left(\mathbf{u}_d^T \mathbf{x}_i - \mathbf{u}_d^T \mathbf{x}_j \right)^2 W_{ij} + \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \left(\mathbf{u}_d^T \mathbf{x}_i + \mathbf{u}_d^T \mathbf{x}_j \right)^2 W_{ij} \right]$$
$$= \frac{1}{2} \sum_{d=1}^{r} \left[\sum_{i,j=1}^{n} \left(\mathbf{u}_d^T \mathbf{x}_i - \mathbf{u}_d^T \mathbf{x}_j \right)^2 W_{ij} + 4 \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \left(\mathbf{u}_d^T \mathbf{x}_i \right) \left(\mathbf{u}_d^T \mathbf{x}_j \right) W_{ij} \right]$$
$$= \frac{1}{2} \sum_{d=1}^{r} \left[2 \sum_{i=1}^{n} \left(\mathbf{u}_d^T \mathbf{x}_i \right)^2 D_{ii} - 2 \sum_{i,j=1}^{n} \left(\mathbf{u}_d^T \mathbf{x}_i \right) \left(\mathbf{u}_d^T \mathbf{x}_j \right) W_{ij} + 4 \sum_{i,j=1}^{n} \left(\mathbf{u}_d^T \mathbf{x}_i \right) \left(\mathbf{u}_d^T \mathbf{x}_j \right) T_{ij} \right]$$
$$= \sum_{d=1}^{r} \mathbf{u}_d^T X(D - W + 2T)X^T \mathbf{u}_d = \sum_{d=1}^{r} \mathbf{u}_d^T XMX^T \mathbf{u}_d = \mathrm{tr}(U^T XMX^T U), \tag{8}$$

where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal elements equal the sums of the row entries of W, i.e., $D_{ii} = \sum_{j=1}^{n} W_{ij}$, $M = D - W + 2T \in \mathbb{R}^{n \times n}$, and $\mathrm{tr}(\cdot)$

stands for the *trace* operator. Note that $L = D - W$ is well known as the *graph Laplacian*. The matrix $M = L + 2T$ is thus the combination of graph Laplacian matrix L and dissimilarity matrix T.

Importantly, the regularizer $\mathcal{R}(U) = \mathrm{tr}(U^T X M X^T U)$ in terms of the transform U is convex because the matrix XMX^T is positive semidefinite ($XMX^T \succeq 0$ has been proved in Eqn. (3.3) as $\mathcal{R}(U) \geq 0$ for any U). Previous metric learning methods [5][7] treat the dissimilarity side information as hard constraints, but we leverage the dissimilarity constraints into the convex regularizer, which sheds light on efficient optimization. We call the formulated regularizer $\mathcal{R}(U) = \mathrm{tr}(U^T X M X^T U)$ as the *dissimilarity-enhanced regularizer* since the core matrix M engages the dissimilarity information other than the similarity information. Our regularizer is similar to the label regularizer proposed in [13] in utilizing dissimilarity information.

3.4 Regularization Framework

Without loss of generality, we suppose the first l samples in X are involved in the side information and form $X_l = [\mathbf{x}_1, \cdots, \mathbf{x}_l] \in \mathbb{R}^{m \times l}$. Using the above regularizer, we propose a novel distance metric learning approach, called *Output Regularized Metric Learning* (ORML), based on the following regularization framework

$$\min_{U \in \mathbb{R}^{m \times r}} \quad \mathrm{tr}(U^T X M X^T U) + \gamma \left\| U^T X_l - Y_l \right\|_{\mathrm{F}}^2 \tag{9}$$

$$s.t. \quad U^T U = \Sigma \tag{10}$$

where $Y_l \in \mathbb{R}^{r \times l}$ is the *ideal output* of some conceived linear transform $\widetilde{U}$ applied to the data matrix X_l such that the output $Y_l = \widetilde{U}^T X_l$ perfectly satisfies the pairwise constraints in $\mathcal{S} \cup \mathcal{D}$. The least squares formulation $\|U^T X_l - Y_l\|_{\mathrm{F}}^2$ instantiates the loss function $\mathcal{V}(A, \mathcal{S}, \mathcal{D})$ stated in the regularization prototype Eqn. (3). $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix with positive entries, i.e., $\Sigma \succ 0$. More clearly, the constraint in Eqn. (10) is equivalent to

$$\mathbf{u}_i^T \mathbf{u}_j = 0, \ i, j = 1, \cdots, r, \ i \neq j. \tag{11}$$

It indicates that $U = [\mathbf{u}_1, \cdots, \mathbf{u}_r]$ consists of r orthogonal vectors in $\mathbb{R}^m$. The reason to impose such an orthogonal constraint is to explicitly make the projection vectors $\mathbf{u}_1, \cdots, \mathbf{u}_r$ linearly independent and, more notably, uncorrelated. Actually, $\mathbf{u}_1, \cdots, \mathbf{u}_r$ are principle eigenvectors of the metric matrix A and are thus physically meaningful to construct the metric as $A = UU^T = \sum_{i=1}^r \mathbf{u}_i \mathbf{u}_i^T$.

The major merit of the proposed regularization framework Eqn. (9)(10) is to adroitly drop the positive semidefinite constraint $A \succeq 0$ in Eqn. (3) which casts the metric learning problem into Semidefinite Programming (SDP) [14]. SDP takes an expensive optimization cost and even becomes computationally prohibitive when the dimension of A is large, e.g., $m > 10^3$. Equivalently, we optimize the transformation matrix U instead of the metric matrix A and thus formulate the metric learning task as a constrained quadratic optimization problem which can be solved quite efficiently with a global optimum solution. In the

next section, we will show the skills for finding the ideal output Y_l as well as coping with the orthogonal constraint in Eqn. (11).

4 ORML Algorithm for CIR

Now, we discuss how to apply ORML to collaborative image retrieval (CIR) and implement related optimization in details. As in the previous work in [1,10], we assume the log data are collected in the form of *log sessions*, of which each one corresponds to a particular user querying process. During each log session, a user first submits an image example to a CBIR system and then judges the relevance on the top ranked images returned by the CBIR system. The relevance judgements specified by the user and the involved image samples, i.e., log samples, are then saved as the log data.

Within each log session of the user's relevance feedback, we can convert the relevance judgements to similar and dissimilar pairwise constraints. For instance, given the query image $\mathbf{x}_i$ and each top-ranked image $\mathbf{x}_j$, if they are marked as relevant in one log session q, we will put $(\mathbf{x}_i, \mathbf{x}_j)$ into the set of similar pairwise constraints $\mathcal{S}_q$; if they are marked as irrelevant, we will put $(\mathbf{x}_i, \mathbf{x}_j)$ into the set of dissimilar pairwise constraints $\mathcal{D}_q$. Note that the first element $\mathbf{x}_i$ in an ordinal pair $(\mathbf{x}_i, \mathbf{x}_j)$ always represents a query image. Consequently, we denote the collection of log data as $\{(\mathcal{S}_q \cup \mathcal{D}_q)|q = 1, \ldots, Q\}$, where Q is the total number of log sessions. The log data exactly provide the side information needed by distance metric learning.

4.1 Ideal Output

Eqn. (9) is essentially a quadratically constrained quadratic optimization problem, and is not easy to solve directly. Here we adopt a heuristic method to explore the solution.

First, we can get an initial transformation matrix V with Principal Component Analysis (PCA) [2]. Without loss of generality, we assume that $\{\mathbf{x}_i\}_{i=1}^n$ be zero-centered. This can be simply achieved by subtracting the mean vector from all $\mathbf{x}_i$s. Let P contain $r \leq \min\{m, n\}$ unitary eigenvectors of XX^T, i.e., $P = [\mathbf{p}_1, \cdots, \mathbf{p}_r]$, corresponding to the r largest eigenvalues $\lambda_1, \cdots, \lambda_r$ with a nonincreasing order. We define the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_r)$ and have $P^T XX^T P = \Lambda$. Then we acquire the initial transform $V \in \mathbb{R}^{m \times r}$ by

$$V = P\Lambda^{-1/2}, \tag{12}$$

such that $V^T XX^T V = \Lambda^{-1/2} P^T XX^T P \Lambda^{-1/2} = I$. For any column vector $\mathbf{v} \in \mathbb{R}^m$ in V and any two inputs $\mathbf{x}_i$ and $\mathbf{x}_j$, we utilize $\sum_{i=1}^n \left(\mathbf{v}^T \mathbf{x}_i\right)^2 = \mathbf{v}^T XX^T \mathbf{v} = 1$ to conclude

$$|\mathbf{v}^T \mathbf{x}_i - \mathbf{v}^T \mathbf{x}_j| = \sqrt{(\mathbf{v}^T \mathbf{x}_i - \mathbf{v}^T \mathbf{x}_j)^2} \leq \sqrt{2\left((\mathbf{v}^T \mathbf{x}_i)^2 + (\mathbf{v}^T \mathbf{x}_j)^2\right)} \leq \sqrt{2}, \tag{13}$$

which indicates that the range of 1D projections $\{\mathbf{v}^T \mathbf{x}_i\}$ on vector $\mathbf{v}$ is upper-bounded.

Let us suppose that the pairwise constraints are imposed on l log samples $\{\mathbf{x}_1, \cdots, \mathbf{x}_l\}$. Thus, we only need to find the output Y_l of X_l. In light of Eqn. (13), we may correct the output $V^T X_l$ under the initial transform V piloted by the constraints $\mathcal{S}_q \cup \mathcal{D}_q$ within each log session q.

Concretely, we investigate each row $\mathbf{v}_d^T X_l$ of $V^T X_l$ and form each row vector $\mathbf{y}^{(d)} \in \mathbb{R}^l$ in output Y_l as follows ($d = 1, \cdots, r$)

$$y_j^{(d)} = \begin{cases} \mathbf{v}_d^T \mathbf{x}_i, & (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_q \\ -\mathrm{sgn}(\mathbf{v}_d^T \mathbf{x}_i)(|\mathbf{v}_d^T \mathbf{x}_i| + \frac{1}{\sqrt{r}}), & (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}_q \end{cases} \tag{14}$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function, returning -1 for negative input and 1 otherwise. The idea of setting $y_j^{(d)}$ based on the PCA output $\mathbf{v}_d^T \mathbf{x}_j$ and side information $\mathcal{S}_q \cup \mathcal{D}_q$ is in tune with the proposed regularizer in Eqn. (3.3) as it turns out that $y_i^{(d)} - y_j^{(d)} = 0$, $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_q$ and $|y_i^{(d)} + y_j^{(d)}| = \frac{1}{\sqrt{r}}$, $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}_q$. The residue $1/\sqrt{r} < 1$ prevents the freak case $y_i^{(d)} = y_j^{(d)} = 0$, $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}_q$.

Throughout all log sessions ($q = 1, \cdots, Q$), we sequentially set up $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(r)}$ using Eqn. (14) and ultimately arrive at

$$Y_l = [\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(r)}]^T = [\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_l] \in \mathbb{R}^{r \times l}, \tag{15}$$

in which $\widehat{\mathbf{x}}_i$ is the low-dimensional representation of $\mathbf{x}_i$ via some conceived linear mapping $\widetilde{U} : \mathbf{x}_i \mapsto \widehat{\mathbf{x}}_i = \widetilde{U}^T \mathbf{x}_i$. Importantly, Y_l exactly obeys all those pairwise constraints $\{\mathcal{S}_q \cup \mathcal{D}_q\}$ because we have

$$\|\widehat{\mathbf{x}}_i - \widehat{\mathbf{x}}_j\| = 0,\ (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_q;\ \|\widehat{\mathbf{x}}_i - \widehat{\mathbf{x}}_j\| \geq 1,\ (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}_q, \tag{16}$$

which implies that the distances between ideal low-dimensional points with similar constraints are zeros and the distances between those with dissimilar constraints are always larger than a constant.

In summary, the found output Y_l perfectly conforms to the min-max principle with skillfully modifying the heuristic output $V^T X_l$ supported by PCA.

4.2 Orthogonal Pursuit

If the constraint in Eqn. (10) is removed, Eqn. (9) can be easily solved and even result in a close-form solution. Eqn. (9) is rewritten as

$$\begin{aligned} & \mathrm{tr}(U^T X M X^T U) + \gamma \left\| U^T X_l - Y_l \right\|_{\mathrm{F}}^2 \\ = & \sum_{d=1}^{r} \mathbf{u}_d^T X M X^T \mathbf{u}_d + \gamma \sum_{d=1}^{r} \left\| X_l^T \mathbf{u}_d - \mathbf{y}^{(d)} \right\|^2 \\ = & \sum_{d=1}^{r} (\mathbf{u}_d^T X M X^T \mathbf{u}_d + \gamma \left\| X_l^T \mathbf{u}_d - \mathbf{y}^{(d)} \right\|^2), \end{aligned} \tag{17}$$

which guides us to greedily pursue the target vectors $\mathbf{u}_d$ ($d = 1, \cdots, r$).

Now we tackle the orthogonal constraint with a recursive notion. Suppose we have obtained $d-1$ $(1 \leq d \leq r)$ orthogonal projection vectors $\mathbf{u}_1, \cdots, \mathbf{u}_{d-1}$, and calculate

$$V^{(d)} = \prod_{i=1}^{d-1} \left(I - \frac{\mathbf{u}_i \mathbf{u}_i^T}{\|\mathbf{u}_i\|^2} \right) [\mathbf{v}_d, \cdots, \mathbf{v}_r] \in \mathbb{R}^{m \times (r-d+1)}, \tag{18}$$

where $V^{(1)} = V = [\mathbf{v}_1, \cdots, \mathbf{v}_r]$.

We constrain $\mathbf{u}_d$ to be the form of $V^{(d)}\mathbf{b}$ ($\mathbf{b}$ is an arbitrary vector) and have the following proposition.

Proposition. $\mathbf{u}_d = V^{(d)}\mathbf{b}$ *is orthogonal to previous* $d-1$ *vectors* $\{\mathbf{u}_1, \cdots, \mathbf{u}_{d-1}\}$.

Proof. For $1 \leq i \leq d-1$, we have

$$\begin{aligned} \mathbf{u}_i^T \mathbf{u}_d = \ & \mathbf{u}_i^T V^{(d)} \mathbf{b} = \mathbf{u}_i^T \left(I - \frac{\mathbf{u}_i \mathbf{u}_i^T}{\|\mathbf{u}_i\|^2} \right) \prod_{t \neq i}^{d-1} \left(I - \frac{\mathbf{u}_t \mathbf{u}_t^T}{\|\mathbf{u}_t\|^2} \right) [\mathbf{v}_d, \cdots, \mathbf{v}_r] \mathbf{b} \\ = \ & (\mathbf{u}_i^T - \mathbf{u}_i^T) \prod_{t \neq i}^{d-1} \left(I - \frac{\mathbf{u}_t \mathbf{u}_t^T}{\|\mathbf{u}_t\|^2} \right) [\mathbf{v}_d, \cdots, \mathbf{v}_r] \mathbf{b} = 0. \end{aligned} \tag{19}$$

□

This proposition shows that the expression $\mathbf{u}_d = V^{(d)}\mathbf{b}$ must satisfy the orthogonal constraint. To obtain the exact solution, we substitute $\mathbf{u}_d = V^{(d)}\mathbf{b}$ into Eqn. (17) and derive

$$\min_{\mathbf{b}} \ \mathbf{b}^T {V^{(d)}}^T X M X^T V^{(d)} \mathbf{b} + \gamma \left\| X_l^T V^{(d)} \mathbf{b} - \mathbf{y}^{(d)} \right\|^2, \tag{20}$$

whose derivatives with respect to $\mathbf{b}$ will vanish at the minimizer $\mathbf{b}^*$. In the sequel, we get the close-form solution for each projection vector:

$$\mathbf{u}_d = V^{(d)} \mathbf{b}^* = V^{(d)} \left[{V^{(d)}}^T \left(\frac{1}{\gamma} X M X^T + X_l X_l^T \right) V^{(d)} \right]^{-1} {V^{(d)}}^T X_l \mathbf{y}^{(d)}. \tag{21}$$

4.3 Algorithm

We summarize the Output Regularized Metric Learning (ORML) algorithm for CIR below. It is appreciable that the number r of the learnt projection vectors is independent of the size l of log samples and can stretch until $r = m$ ($m < n$ in this paper).

1. **Compute the regularizer:** Build two weight matrices W and T upon all n input samples with Eqn. (4) and Eqn. (5), respectively. Calculate the graph Laplacian matrix L and the matrix $M = L + 2T$. Compute the matrix $S = XMX^T \in \mathbb{R}^{m \times m}$ used in the dissimilarity-enhanced regularizer.

2. **PCA:** Run PCA on $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \in \mathbb{R}^m$ to get the matrix $V = [\mathbf{v}_1, \cdots, \mathbf{v}_r] \in \mathbb{R}^{m \times r}$ $(r \leq m)$ such that $V^T X X^T V = I$.

3. **Get output:** Given the log data $X_l \in \mathbb{R}^{m \times l}$ and $\{\mathcal{S}_q, \mathcal{D}_q\}_{q=1}^{Q}$, use Eqn. (14) and Eqn. (15) to get the output matrix $Y_l = [\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(r)}]^T \in \mathbb{R}^{r \times l}$.

4. **Orthogonal pursuit:**
   For $d = 1$ to r
   $\mathbf{u}_d \longleftarrow V\left[V^T\left(\frac{1}{\gamma}S + X_l X_l^T\right)V\right]^{-1} V^T X_l \mathbf{y}^{(d)}$;
   $V \longleftarrow$ exclude the first column of V;
   $V \longleftarrow \left(I - \frac{\mathbf{u}_d \mathbf{u}_d^T}{\|\mathbf{u}_d\|^2}\right)V$;
   End.
5. **Construct the metric:** Form the projection matrix $U = [\mathbf{u}_1, \cdots, \mathbf{u}_r] \in \mathbb{R}^{m \times r}$, and then construct the distance metric as $A = UU^T$.

5 Experiments

In our experiments, we evaluate the effectiveness of the proposed ORML algorithm applied to collaborative image retrieval. We design the experiments in several perspectives for extensive performance evaluation, where we compare ORML with state-of-the-art distance metric learning techniques through engaging normal, limited and noisy log data.

We obtained a standard CBIR testbed from the authors in [1]. The testbed consists of real-world images from COREL image CDs. It contains two datasets : 20-Category (20-Cat) that includes images from 20 different categories, and 50-Category (50-Cat) that includes images from 50 different categories. Each category consists of exactly 100 images that are randomly selected from relevant examples in the COREL CDs. Every category represents a different semantic topic, such as *antelope*, *balloon*, *butterfly*, *car*, *cat*, *dog*, *horse*, etc. The way of sampling the images with semantic categories lets us evaluate the retrieval performance automatically, which significantly reduces the subjective errors caused by manual evaluations. In this paper, we only employ 50-Cat since it provides us much more samples.

5.1 Image Representation and Log Data

We use color, edge and texture to represent images. Three types of color moments, mean, variance and skewness, are extracted in each color channel (H, S, and V), resulting in a 9-dimensional color moment feature vector. The Canny edge detector is applied to obtain the edge image from which the edge direction histogram is computed. Each edge direction histogram is quantized into 18 bins of 20 degrees each and an 18-dimensional edge feature vector is acquired. To extract texture features, the Discrete Wavelet Transformation (DWT) is performed on the image using a Daubechies-4 wavelet filter [15]. For each image, 3-level DWT decompositions are conducted and the entropy values of 9 resulting subimages are computed, which gives rise to a 9-dimensional texture feature vector. Finally, an image is represented as a 36-dimensional feature vector.

We collected the real log data related to the COREL testbed by a real CBIR system from the authors in [1]. In their collection, there are two sets of log data. One is a set of normal log data, which contains small noise. The other is a set of noisy log data with relatively large noise. In log data, a log session is defined as

Table 1. The log data collected from users

Dataset	Normal Log		Noisy Log	
	#Log Sessions	Noise	# Log Sessions	Noise
50-Cat	150	7.7%	150	17.1%

the basic unit. Each log session corresponds to a customary relevance feedback process, in which 20 images were judged by a user. So, each log session contains 20 log samples that are marked as either relevant or irrelevant. Table 1 shows the basic information of the log data.

5.2 Experimental Setup

We compare ORML with representative metric learning techniques using side information. It is noticeable that we do not compare ORML with supervised metric learning approaches as they require explicit class labels for classification, which is unsuitable for CIR. The compared approaches include six distance metric techniques:

- **Euclidean**: The baseline denoted as "EU".
- **Xing**: A pairwise constraint-based metric learning method with an iterative convex optimization procedure [7].
- **RCA**: Relevance Component Analysis, which learns linear transformations using only equivalent (similar) constraints [8].
- **DCA**: Discriminative Component Analysis, which improves RCA by engaging inequivalent (dissimilar) constraints [9].
- **RML**: Regularized Metric Learning using the Frobenius norm as the regularizer [10].
- **ORML**: the proposed Output Regularized Metric Learning algorithm using the dissimilarity-enhanced regularizer and the output-based loss function.

Lately, an information-theoretic metric learning approach is presented to express the metric learning problem as a Bregman optimization problem [16]. This approach aims at minimizing the differential relative entropy between two multivariate Gaussians under pairwise constraints on the distance function. Due to the space limit in this paper, we do not contrast ORML with the information-theoretic metric learning approach.

We follow the standard procedure in CBIR experiments. Specifically, a query image is picked from the database and then queried with the evaluated six distance metrics. The retrieval performance is evaluated based on top ranked images ranging from top 10 to top 100. The average precision (AP) and Mean Average Precision (MAP) are used as the performance measures, which are broadly adopted in CBIR experiments. To implement ORML, we use $k = 6$ nearest neighbors, $\alpha = 1$ and $\beta = 2$ to compute the matrix M used in our regularizer. The regularization parameter γ is simply fixed to 9 and the number of projection vectors, which construct the target distance metric, is set to $r = 15$.

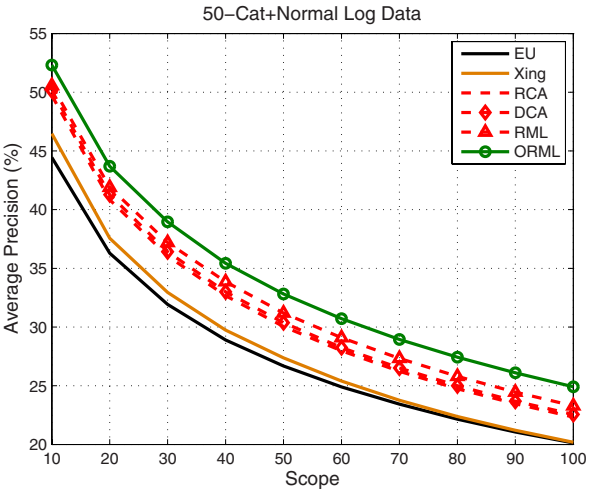

Fig. 1. Average precision of top ranked images on the 50-Category testbed with normal log data

Table 2. Mean Average Precision (%) of top ranked images on the 50-Category testbed over 5,000 queries with three kinds of log data. The last column shows the relative improvement of the proposed ORML over the baseline (Euclidean).

Datasets	EU	Xing	RCA	DCA	RML	ORML	ORML Improve
50-Cat+normal log	27.99	28.71	31.41	31.72	32.47	34.13	21.94
50-Cat+small log	27.99	28.27	31.31	31.64	31.41	32.68	16.76
50-Cat+noisy log	27.99	27.95	29.79	30.14	31.08	33.48	19.61

5.3 Normal Log Data

Above all, we perform experiments on the normal log data. Figure 1 shows the experimental results on the 50-Category dataset. By comparing the four previous metric learning methods, Xing, RCA, DCA and RML, RML achieves the best overall performance, obtaining 16.0% improvement on MAP over the baseline, as shown in Table 2. Xing performs the worst among the four methods. Overall, we find that the proposed ORML algorithm achieves the best performance, significantly improving the baseline by about 21% on MAP. It demonstrates that ORML is more effective than the previous methods when working on the normal log data. Here, we find that Xing *et al.*' method does not perform well in this dataset. One possible reason is that this method may be too sensitive to noise since it imposes the hard constraint on dissimilar data points.

5.4 Small Log Data

In this experiment, we evaluate the robustness of the metric learning methods on small amount of normal log data. This situation usually happens at the beginning

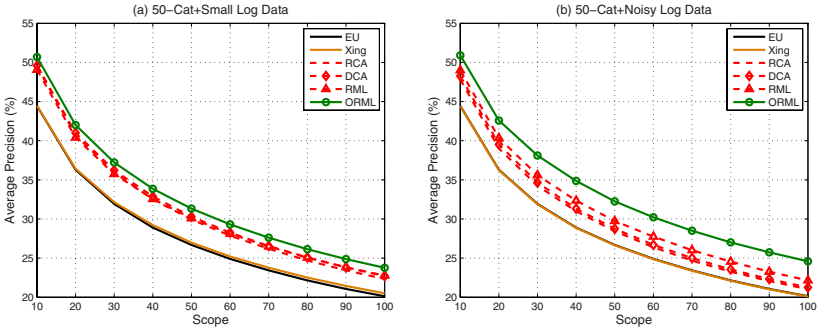

Fig. 2. Average precision of top ranked images on the 50-Category testbed with small and noisy log data.

stage of CBIR systems. Figure 2(a) shows the experimental results on the 50-Category testbed with a small subset of the normal log data which contains only 50 log sessions randomly selected from the normal log dataset. Again, we find that ORML achieves the best improvement among all compared methods over the baseline. It also shows that the proposed ORML method is more effective to learn robust metrics by utilizing the unlabeled data, even with limited log data.

5.5 Noisy Log Data

To further validate the robustness performance, the third experiment is to evaluate the metric learning methods on the noisy log data carrying relatively large noise. Figure 2(b) shows the experimental results on the 50-Category testbed with the noisy log data. From the experimental results, we find that Xing *et al.*'s method fails to improve over the baseline due to the noise problem. The results also validate our previous conjecture that Xing *et al.*'s method may be too sensitive to noise. Compared with Xing, the other three metric learning methods including RCA, DCA and RML are less sensitive to noise, but they still suffer a lot from the noise. For example, RCA achieves 12.2% improvement on MAP with the normal log data as shown in Table 2, but only achieves 6.4% improvement on MAP with the same amount of noisy log data. In contrast, ORML achieves 21.94% improvement on MAP with normal log data, and is able to keep 19.61% improvement on MAP with the larger noisy log data without too much dropping. These experimental results again validate that ORML is effective to learn reliable distance metrics with real noisy log data.

6 Conclusion

This paper studies distance metric learning with side information and its application to collaborative image retrieval, in which real log data provided by the user's relevance feedback are leveraged to improve traditional CBIR performance. To robustly exploit the log data and smoothly incorporate the unlabeled

data, we propose the Output Regularized Metric Learning (ORML) algorithm. ORML uses the dissimilarity-enhanced regularizer and the ideal output of log samples piloted by side information for learning a series of orthogonal projection vectors that readily construct an effective metric. The promising experimental results show that the proposed ORML algorithm is more effective than the state-of-the-arts in learning reliable metrics with real log data.

Acknowledgement

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong SAR, China (Project No. CUHK 414306).

References

1. Hoi, S.C., Lyu, M.R., Jin, R.: A unified log-based relevance feedback scheme for image retrieval. IEEE Trans. Knowledge and Data Engineering 18(4), 509–524 (2006)
2. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Elsevier, Amsterdam (1990)
3. Goldberger, G.H.J., Roweis, S., Salakhutdinov, R.: Neighbourhood components analysis. In: NIPS 17 (2005)
4. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: NIPS 18 (2006)
5. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: NIPS 18 (2006)
6. Yang, L., Jin, R., Sukthankar, R., Liu, Y.: An efficient algorithm for local distance metric learning. In: Proc. AAAI (2006)
7. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side-information. In: NIPS 15 (2003)
8. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning a mahalanobis metric from equivalence constraints. JMLR 6, 937–965 (2005)
9. Hoi, S.C., Liu, W., Lyu, M.R., Ma, W.-Y.: Learning distance metrics with contextual constraints for image retrieval. In: Proc. CVPR (2006)
10. Si, L., Jin, R., Hoi, S.C., Lyu, M.R.: Collaborative image retrieval via regularized metric learning. ACM Multimedia Systems Journal 12(1), 34–44 (2006)
11. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. PAMI 22(12), 1349–1380 (2000)
12. Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons, Chichester (1998)
13. Goldberg, A., Zhu, X., Wright, S.: Dissimilarity in graph-based semi-supervised classification. In: Proc. Artificial Intelligence and Statistics (2007)
14. Boyd, S., Vandenberge, L.: Convex Optimization. Cambridge University Press, Cambridge (2003)
15. Manjunath, B., Newsam, P.W.S., Shin, H.: A texture descriptor for browsing and similarity retrieval. Signal Processing Image Communication (2001)
16. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proc. ICML (2007)