Short Papers

A-Optimal Projection for Image Representation

Xiaofei He, *Senior Member, IEEE*, Chiyuan Zhang, Lijun Zhang, *Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

Abstract—We consider the problem of image representation from the perspective of statistical design. Recent studies have shown that images are possibly sampled from a low dimensional manifold despite of the fact that the ambient space is usually very high dimensional. Learning low dimensional image representations is crucial for many image processing tasks such as recognition and retrieval. Most of the existing approaches for learning low dimensional representations, such as principal component analysis (PCA) and locality preserving projections (LPP), aim at discovering the geometrical or discriminant structures in the data. In this paper, we take a different perspective from statistical experimental design, and propose a novel dimensionality reduction algorithm called A-Optimal Projection (AOP). AOP is based on a linear regression model. Specifically, AOP finds the optimal basis functions so that the expected prediction error of the regression model can be minimized if the new representations are used for training the model. Experimental results suggest that the proposed approach provides a better representation and achieves higher accuracy in image retrieval.

Index Terms—Dimensionality reduction, optimal design, image representation

1 Introduction

IMAGE representation has been a fundamental problem for efficient and effective classification [1], [2], [3] clustering [4], [5], and retrieval [6], [7], [8], [9]. Visual features, such as color, texture, shape, are usually extracted to represent the image. However, the low level feature space is usually of very high dimensionality. Thus, various techniques have been developed for reducing the dimensionality of the feature space, in the hope of obtaining a more manageable problem. If the image space is a linearly embedded manifold, PCA is guaranteed to uncover the intrinsic dimensionality of the manifold and produces a compact representation. However, a number of research efforts have shown that the images possibly reside on a nonlinear submanifold [6], [10]. Particularly, manifold learning techniques, such as Isomap [11], Locally Linear Embedding [12], [13], and Laplacian Eigenmap [14] are proposed to discover the nonlinear structure of the manifold.

All the aforementioned methods try to discover either geometrical or cluster structure hidden in the data. However, they are not directly related to the learning task such as regression. In this paper, we propose a novel dimensionality reduction algorithm called *A-Optimal Projection*, which aims to improve the regression performance in the reduced space. Our approach is motivated by the recent progress on manifold regularized regression, i.e., Laplacian Regularized Least Squares (LapRLS, [15]). Specifically, the loss function of the regression model imposes a locality preserving regularizer into the standard least-square-error based loss function. It finds a classifier which is locally as smooth as possible. Using techniques from optimal experimental design (OED, [16]), we first compute the parameter covariance matrix of the regression model in the reduced space. Then the best projections are learned such that the model parameter variances and the expected prediction error can be minimized. In this work, we adopt A-optimality, which minimizes the trace of the parameter covariance matrix, so the obtained projections are called A-optimal. We further introduce two optimization schemes to solve the objective function. One is based on Semi-Definite Programming (SDP) [17] and the other is based on iterative updating.

It is worthwhile to highlight several aspects of the proposed approach here:

- Similar to many existing manifold learning algorithms, our approach explicitly considers the manifold structure by using a locality preserving regularizer. Our approach can be performed under either unsupervised, supervised, or semi-supervised mode. When there is label information available, it can be naturally encoded in the Laplacian matrix.
- Most of the existing dimensionality reduction algorithms are applied as pre-processing of the data. Similar to [18], [19], [20], our approach takes a different perspective to directly improve the performance of a regularized regression model in the reduced space. Specifically, the model's parameter covariance matrix is minimized in the reduced space, so that the learned regression model can be as stable as possible.
- In this work we adopt the A-optimality to measure the size of the parameter covariance matrix in the reduced space. However, one can also use other optimality criteria such as D-optimality and E-optimality.

2 Related Work

In this section, we give a brief review of related work.

2.1 Linear Dimensionality Reduction

Recently, graph based dimensionality reduction [10], [21], [22] has received considerable interest due to its effectiveness and flexibility. Given a graph G with m vertices, each vertex represents a data point. Let S be a symmetric $m \times m$ matrix with S_{ij} having the weight of the edge joining vertices i and j. The purpose of graph embedding is to represent each vertex of the graph as a low dimensional vector that preserves similarities between the vertex pairs, where similarity is measured by the edge weight.

Let $\mathbf{y} = (y_1, \ldots, y_m)^T$ be the map from the graph to the real line, and $y_i = \mathbf{w}^T\mathbf{x}_i$ where $\mathbf{w}$ is a transformation vector. The locality preserving projections (LPP, [21]) algorithm finds the optimal $\mathbf{w}$ by minimizing [21]:

$$\begin{aligned} &\min \sum_{i,j=1}^{m} (\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j)^2 S_{ij} \\ &s.t.\ \mathbf{w}^T X D X^T \mathbf{w} = 1, \end{aligned} \tag{1}$$

where D is a diagonal matrix whose entries are column (or row, since S is symmetric) sums of W, $D_{ii} = \sum_j S_{ij}$. The constraint $\mathbf{w}^T X D X^T \mathbf{w} = 1$ removes an arbitrary scaling factor in the embedding.

- *X. He and C. Zhang are with the State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou 310058, China. E-mail: xiaofeihe@cad.zju.edu.cn, pluskid@gmail.com.*
- *L. Zhang is with National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. E-mail: zhanglj@lamda.nju.edu.cn.*
- *X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P.R. China. E-mail: xuelong_li@opt.ac.cn.*

Manuscript received 2 Oct. 2011; revised 13 Jan. 2015; accepted 21 May 2015. Date of publication 31 May 2015; date of current version 8 Apr. 2016. Recommended for acceptance by M. Belkin. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2015.2439252

2.2 Laplacian Regularized Least Squares

We consider a linear model

$$y = \mathbf{w}^T\mathbf{x} + \epsilon, \tag{2}$$

where y is the *observation*, $\mathbf{x} \in \mathbf{R}^n$ is the *independent variable*, $\mathbf{w}$ is the *weight vector*, and ϵ is an unknown error with zero mean. Different observations have errors that are independent, but with equal variances σ^2. We define $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ to be the learner's output given input $\mathbf{x}$ and the weight vector $\mathbf{w}$. Given a set of training points $\{\mathbf{x}_i, y_i\}_{i=1}^{\ell}$, the maximum likelihood estimate for the weight vector, $\widehat{\mathbf{w}}$, is that which minimizes the sum of squared error

$$J_{\text{sse}}(\mathbf{w}) = \sum_{i=1}^{\ell}\left(\mathbf{w}^T\mathbf{x}_i - y_i\right)^2. \tag{3}$$

This problem has a closed form solution given by

$$\widehat{\mathbf{w}} = (XX^T)^{-1}X\mathbf{y}, \tag{4}$$

where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_\ell)$ and $\mathbf{y} = (y_1, \ldots, y_\ell)^T$.

LapRLS extends the ordinary linear regression by incorporating geometrical information into the loss function. It constructs a nearest neighbor graph with weight matrix S to model the geometrical structure of the data manifold. Let $\mathcal{N}_k(\mathbf{x})$ denote the k nearest neighbors of $\mathbf{x}$. There are many choices of the weight matrix S. A simple definition is as follows:

$$S_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i); \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

LapRLS seeks for a function which varies as smooth as possible on the manifold by solving the following problem

$$\min_{\mathbf{w}} \sum_{i=1}^{m}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 + \frac{\lambda_1}{2}\sum_{i,j=1}^{m}(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j)^2 S_{ij} + \lambda_2\|\mathbf{w}\|^2,$$

The solution is given by

$$\widehat{\mathbf{w}} = (XX^T + \lambda_1 XLX^T + \lambda_2 I)^{-1}X\mathbf{y}, \tag{6}$$

where $L = diag(S\mathbf{1}) - S$ is called *graph Laplacian* [23] and $\mathbf{1}$ is a vector of all ones.

2.3 Optimal Experimental Design

Consider the same linear model in (2) and the estimation in (4). By Gaussian-Markov theorem, we know that $\hat{\mathbf{w}} - \mathbf{w}$ has a zero-mean and a covariance matrix given by $\sigma^2 H_{\text{sse}}^{-1}$, where H_{sse} is the Hessian of the sum squared error J_{sse} in (3):

$$H_{\text{sse}} = \left(\frac{\partial^2 J_{\text{sse}}}{\partial \mathbf{w}^2}\right) = \left(\sum_{i=1}^{\ell}\mathbf{x}_i\mathbf{x}_i^T\right) = XX^T.$$

The goal of *Optimal Experimental Design* [16] is to seek an *optimal* distribution of the labeled sample points in the sense that the *size* of covariance matrix of $\hat{\mathbf{w}} - \mathbf{w}$ is minimized. There are three most common scalar measures of the size of the covariance matrix in optimal experimental design:

- D-optimal design: determinant of the matrix;
- A-optimal design: trace of the matrix;
- E-optimal design: maximum eigenvalue of the matrix.

3 A-Optimal Projection

In this section, we introduce our A-Optimal Projection for linear dimensionality reduction.

3.1 Problem Definition

Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ be a $n \times m$ data matrix, where n is the number of features and m is the number of data points. Our goal is to find a transformation matrix $A \in \mathbb{R}^{n\times k}$ that maps these m points to a set of points $\mathbf{y}_1, \ldots, \mathbf{y}_m \in \mathbb{R}^k (k \ll n)$, where $\mathbf{y}_i = A^T\mathbf{x}_i$.

Particularly, we consider the situation of using $\mathbf{y}_i$ to train a linear regression model:

$$z = \boldsymbol{\beta}^T\mathbf{y} + \epsilon, \tag{7}$$

where z is the *observation*, $\mathbf{y}$ is the *independent variable*, $\boldsymbol{\beta}$ is the *weight vector* and ϵ is an unknown error with zero mean and variance σ^2.

3.2 The Objective Function

Consider learning a graph regularized regression model by using the new representations $\mathbf{y}_1, \ldots, \mathbf{y}_m$ and their labels $z_1, \ldots, z_m$:

$$J_{LapRLS}(\boldsymbol{\beta}) = \sum_{i=1}^{m}\left(z_i - \boldsymbol{\beta}^T\mathbf{y}_i\right)^2 + \frac{\lambda_1}{2}\sum_{i,j=1}^{m}\left(\boldsymbol{\beta}^T\mathbf{y}_i - \boldsymbol{\beta}^T\mathbf{y}_j\right)^2 S_{ij} + \lambda_2\|\boldsymbol{\beta}\|^2.$$

It is easy to check that the solution is given by

$$\widehat{\boldsymbol{\beta}} = \left(YY^T + \lambda_1 YLY^T + \lambda_2 I\right)^{-1}Y\mathbf{z}, \tag{8}$$

where L is the Laplacian matrix defined on the graph, $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$ is the new data matrix and $\mathbf{z} = (z_1, \ldots, z_m)$ is the vector of labels. We define

$$H = YY^T + \lambda_1 YLY^T + \lambda_2 I,$$

and

$$\Lambda = \lambda_1 YLY^T + \lambda_2 I.$$

By noticing that $Cov(\mathbf{z}) = \sigma^2 I$ where I is the identity matrix, the covariance matrix of the parameter $\boldsymbol{\beta}$ can be computed as follows:

$$\begin{aligned} Cov(\widehat{\boldsymbol{\beta}}) &= Cov(H^{-1}Y\mathbf{z}) = H^{-1}Y\,Cov(\mathbf{z})Y^TH^{-1} \\ &= \sigma^2 H^{-1}YY^TH^{-1} = \sigma^2 H^{-1}(H - \Lambda)H^{-1} \\ &= \sigma^2(H^{-1} - H^{-1}\Lambda H^{-1}). \end{aligned}$$

Since λ_1 and λ_2 are usually set to be very small, we have

$$Cov(\widehat{\boldsymbol{\beta}}) \approx \sigma^2 H^{-1}. \tag{9}$$

Recall that $Y = A^TX$, we have

$$\begin{aligned} Cov(\widehat{\boldsymbol{\beta}}) &\approx \sigma^2(A^TXX^TA + \lambda_1 A^TXLX^TA + \lambda_2 I)^{-1} \\ &= \sigma^2(A^TX(I + \lambda_1 L)X^TA + \lambda_2 I)^{-1}. \end{aligned} \tag{10}$$

Given a data point $\mathbf{y}$ in the new representation space, its expected prediction error has the expression $\mathbf{y}^T Cov(\widehat{\boldsymbol{\beta}})\mathbf{y}$. In order to minimize the expected prediction error, one has to minimize the size of the parameter covariance matrix.

In statistics, there are many different ways to measure the size of the parameter covariance matrix [16]. In this work, we apply A-optimality [16] which minimizes the trace of the parameter covariance matrix. Thus, the optimal transformation matrix A can be obtained by solving the following objective function:

$$\min_A \text{Tr}\left((A^TX(I + \lambda_1 L)X^TA + \lambda_2 I)^{-1}\right). \tag{11}$$

It is easy to check that the matrix $I + \lambda_1 L$ is symmetric and positive definite. Thus, we can decompose it as follows:

$$I + \lambda_1 L = \Sigma\Sigma^T.$$

We define

$$\widetilde{X} = X\Sigma.$$

The optimization problem (11) can be rewritten as follows:

$$\min_A \mathrm{Tr}\big((A^T\widetilde{X}\widetilde{X}^TA+\lambda_2 I)^{-1}\big). \tag{12}$$

We have the following theorem:

Theorem 3.1. *The optimization problem (12) is equivalent to the following:*

$$\min_A \mathrm{Tr}\big((\widetilde{X}^TAA^T\widetilde{X}+\lambda_2 I)^{-1}\big). \tag{13}$$

The proofs of all the theorems can be found in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2015.2439252.

Notice that for dimensionality reduction, we only care about the direction of the projection, and the scale of the projection is not important. Thus, it is necessary to remove the freedom of scaling in (13), which can be achieved by imposing an constraint that $\|A\|_F \leq \rho$ or adding a regularizer $\gamma\|A\|_F^2$ in the objective function.

3.3 Relations to Other Methods

Although our algorithm is formulated based on a linear model for supervised learning, we can compute the projection matrix in a fully unsupervised manner. When the label information is available, it also can be easily incorporated into our approach. For example, when constructing the Laplacian matrix L, we can use the provided label information to build the weight graph S. Specifically, we connect the points $\mathbf{x}_i$ and $\mathbf{x}_j$ if they share the same label. This is the standard way of incorporating the label information into the geometric characterization of the data manifold [24]. In this way, the projection subspace gets a strong dependency on the regression variable z.

Besides the linear dimensionality reduction algorithms we mentioned in Section 2.1, there are several other related subspace regression algorithms. Among them, the simplest one is *principal component regression* (PCR) [25]. It simply performs the *principal component analysis* (PCA) to project the data to a lower dimensional subspace and then performs linear regression in the subspace. In our algorithm, if we set both of the regularization coefficients λ_1 and λ_2 to be zero in the objective function (11), we get

$$\min_A \mathrm{Tr}((A^TXX^TA)^{-1}). \tag{14}$$

It can be easily seen that this is equivalent to the objective function of PCA.

PCR does not use the label information when computing the subspace. *Partial least squares regression* (PLS) [26] projects both the input variable x and the regression variable z to subspaces of the same dimension by maximizing the covariance between them after projection. Unlike PCR and our method, PLS computes projection and regression simultaneously.

Finally, *sliced inverse regression* (SIR) [27] seeks for a subspace of x that captures all the dependency between the response z and the input x. In other words, it tries to ensure that $z \perp x|y$, where y is the variable for the projected input x. The basic idea of the SIR algorithm is to "reverse" the role and perform regression of the input x on the label z. However, this kind of inverse regression generally requires making assumptions with respect to the probability distribution of the input x, which can be difficult to justify [28]. Our model, on the other hand, tries to minimize the covariance of the model parameters as well as the expected prediction error in the reduced subspace.

4 Optimization

In this section, we introduce two optimization schemes to solve the objective function (13).

4.1 Convex Optimization

In this section, we describe how to solve the optimization problem (13) by using semi-definite programming [17], [29].

It is well-known that AA^T is symmetric and positive semi-definite. On the other hand, for any symmetric and positive semi-definite matrix, it can be decomposed to the form of AA^T. Let $M = AA^T$. Let $\mathbb{S}_n^+$ denote the set of symmetric and positive semi-definite $n \times n$ matrices. The associated generalized inequality $\succeq_{\mathbb{S}_n^+}$ is the usual matrix inequality: $A \succeq_{\mathbb{S}_n^+} B$ means $A - B$ is a positive semi-definite matrix. Thus, the optimization problem (13) can be reduced to

$$\begin{aligned} &\min \ \ \mathrm{Tr}((\lambda I+\widetilde{X}^TM\widetilde{X})^{-1}),\\ &s.t. \ \ M \succeq_{\mathbb{S}_n^+} 0. \end{aligned} \tag{15}$$

The following theorem shows that the optimization problem (15) is convex with variable $M \in \mathbb{R}^{n\times n}$.

Theorem 4.1. *The optimization problem (15) is convex with variable* $M \in \mathbb{R}^{n\times n}$.

By introducing a new variable $P \in \mathbb{R}^{n\times n}$, the optimization problem (15) can be equivalently rewritten as follows:

$$\begin{aligned} &\min \ \ \mathrm{Tr}(P)\\ &s.t. \ \ P \succeq_{\mathbb{S}_n^+} (\lambda I+\widetilde{X}^TM\widetilde{X})^{-1}\\ &\quad\ \ M \succeq_{\mathbb{S}_n^+} 0 \end{aligned} \tag{16}$$

with variables P and M.

Theorem 4.2. *The optimization problem (16) is equivalent to the optimization problem (15).*

In the following, we discuss how to use Schur complement theory [17] to cast the optimization problem (16) as a semi-definite programming. Suppose A, B, C are respectively $p \times p$, $p \times q$ and $q \times q$ matrices, and A is invertible. Let

$$Q = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}.$$

The Schur complement of the block A of the matrix Q is the $p \times p$ matrix $C - B^TA^{-1}B$. Schur complement theorem states that Q is positive semi-definite if and only if $C - B^TA^{-1}B$ is positive semi-definite [17]. By using Schur complement theorem, the optimization problem (16) can be expressed as

$$\begin{aligned} &\min \ \ \mathrm{Tr}(P)\\ &s.t. \ \ \begin{pmatrix} \lambda I+\widetilde{X}^TM\widetilde{X} & I \\ I & P \end{pmatrix} \succeq_{\mathbb{S}_n^+} 0\\ &\quad\ \ M \succeq_{\mathbb{S}_n^+} 0. \end{aligned} \tag{17}$$

Similar to the discussion in the end of Section 3.2, we also need to control the size of M. Since $M = AA^T$ and $\|A\|_F^2 = \mathrm{Tr}(AA^T)$, we can impose an constraint that $\mathrm{Tr}(M) \leq \rho$ or add a regularizer $\gamma\mathrm{Tr}(M)$.

The above optimization problem can be solved by using interior-point methods [17]. Once we obtain M, we can compute the eigen-decomposition of M, which expresses

$$M = UVU^T, \tag{18}$$

where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ is an $n \times n$ orthonormal matrix, i.e., $UU^T = U^TU = I$, and $V = diag(v_1, \ldots, v_n)$ is a diagonal matrix. The column vectors of U are the eigenvectors of M and the diagonal entries of V are the corresponding eigenvalues of M. Without loss

of generality, we assume $v_1 \geq \cdots \geq v_n \geq 0$. Thus, the optimal transformation matrix A is given by

$$A = (\mathbf{a}_1, \ldots, \mathbf{a}_k), \ \mathbf{a}_i = \sqrt{v_i}\mathbf{u}_i. \tag{19}$$

4.2 Iterative Optimization

Although the optimization scheme described above is guaranteed to find the global optimum, SDP is computationally very expensive and thus may not be applicable to real world applications. In this section, we discuss an alternative optimization scheme which is much more efficient.

We have the following theorem:

Theorem 4.3. *The optimization problem (13) is equivalent to the following optimization problem:*

$$\min_{A,B} \|I - A^T\widetilde{X}B\|^2 + \lambda\|B\|^2, \tag{20}$$

where $A \in \mathbb{R}^{n\times k}$ and $B \in \mathbb{R}^{m\times k}$ are the two variables.

Theorem 4.3 tells us that the optimal A can be obtained by iteratively computing A and B. Suppose B is given. Denote the objective function in (20) by ϕ. Suppose B is given. By requiring the gradient of ϕ with respect to B^T vanish, we have

$$\begin{aligned} &\frac{\partial \phi}{\partial B^T} = 0 \\ &\quad \Rightarrow B^T\widetilde{X}^T AA^T\widetilde{X} + B^T - A^T\widetilde{X} = 0 \\ &\quad \Rightarrow B = (\widetilde{X}^T AA^T\widetilde{X} + \lambda I)^{-1}\widetilde{X}^T A. \end{aligned} \tag{21}$$

Suppose B is given. In order to compute A, we take the partial derivative of ϕ with respect to A and require it vanish:

$$\begin{aligned} &\frac{\partial \phi}{\partial A} = 0 \\ &\quad \Rightarrow -2\frac{\partial \mathrm{Tr}(A^T\widetilde{X}B)}{\partial A} + \frac{\partial \mathrm{Tr}(A^T\widetilde{X}BB^T\widetilde{X}^T A)}{\partial A} = 0 \\ &\quad \Rightarrow -2\widetilde{X}B + 2\widetilde{X}BB^T\widetilde{X}^T A = 0 \\ &\quad \Rightarrow A = (\widetilde{X}BB^T\widetilde{X})^{-1}\widetilde{X}B. \end{aligned} \tag{22}$$

The algorithmic procedure of computing the projection matrix A can be summarized as follows:

1) Initialize the matrix A by computing the PCA of the data, that is, the principal eigenvectors of the covariance matrix;
2) Compute the matrix B according to Eq. (21);
3) If we control the size of A via an constraint $\|A\|_F \leq \rho$
   - Update the matrix A according to Eq. (22),
   - Normalize A to satisfy this constraint;

   else if we control the size of A via a regularizer $\gamma\|A\|_F^2$
   - Update the matrix A according to

$$A = (\widetilde{X}BB^T\widetilde{X} + \gamma I)^{-1}\widetilde{X}B.$$

4) Repeat steps 2 and 3 until convergence.

5 Experimental Results

In this section, we evaluate the performance of our proposed approach on image retrieval. We apply different dimensionality reduction algorithms to project the images into a lower dimensional subspace in which LapRLS is applied to rank the images.

5.1 Relevance Feedback Image Retrieval

Relevance feedback is a well established and effective framework for narrowing down the gap between low-level visual features and high-level semantic concepts in content-based image retrieval (CBIR) [30]. In this experiments, we compare the following state-of-the-art algorithms:

- Locality preserving projections (**LPP**, [21]), as described in Section 2.1.
- Augmented relational embedding (**ARE**, [10]). Unlike LPP [21], ARE uses an additional graph to encode the label information provided by user's relevance feedbacks.
- Semantic subspace projection (**SSP**, [22]). Similar to ARE, SSP also uses an additional graph to encode the label information.
- Our proposed **AOP** algorithm. In our algorithm, the label information is incorporated into the weight matrix by assigning a higher weight to the image pair from the same semantic class:

$$S_{ij}^{AOP} = \begin{cases} \alpha, & \text{if } \mathbf{x}_i \in \boldsymbol{F}^+ \text{ and } \mathbf{x}_j \in \boldsymbol{F}^+; \\ 1, & \text{if } \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i); \\ 0, & \text{otherwise,} \end{cases} \tag{23}$$

  where α is a suitable constant. Notice that the Laplacian matrix is computed before dimensionality reduction and used in both AOP and LapRLS. For optimization, we choose the iterative scheme developed in Section 4.2, and we use a small regularizer to control the norm of A.

5.2 Data Preparation

The image database consists of 5,000 images of 50 semantic categories from the COREL data set. It is a large and heterogeneous image set. In this work, we combine a 64-dimensional color histogram and a 64-dimensional color texture moment (CTM) [31] to represent the images. The color histogram is calculated using $4 \times 4 \times 4$ bins in HSV space. CTM, which was proposed by Yu et al. [31], integrates the color and texture characteristics of the image in a compact form. CTM adopts local Fourier transform as a texture representation scheme and derives eight characteristic maps for describing different aspects of co-occurrence relations of image pixels in each channel of the (SVcosH, SVsinH, and V) color space. Then, CTM calculates the first and second moments of these maps as a representation of the natural color image pixel distribution. Please see [31] for details.

5.3 Evaluation Settings

We use *precision-scope curve* and *precision rate* [32] to evaluate the effectiveness of the image retrieval algorithms. The scope is specified by the number N of top-ranked images presented to the user. The precision is the ratio of the number of relevant images presented to the user to the scope N. The precision-scope curve describes the precision with various scopes and thus gives the overall performance evaluation of the algorithms. On the other hand, the precision rate emphasizes the precision at a particular value of scope. In general, it is appropriate to present 20 images on a screen, and thus the precision at the top 20 is especially important.

We perform *five fold cross validation* to evaluate the algorithms. More precisely, we divide the whole image database into five subsets with equal size. Thus, there are 20 images per category in each subset. At each run of cross validation, one subset is selected as the query set, and the other four subsets are used as the database for retrieval. The precision-scope curve and precision rate are computed by averaging the results from the five-fold cross validation.

We designed an automatic feedback scheme to model the retrieval process. For each submitted query, our system retrieves and ranks the images in the database. The top 10 ranked images were selected as the feedback images, and their label information

TABLE 1
Precision at Top 20 Returns of the 4 Algorithms after the 1st Feedback Iteration (Mean%)

Category	Baseline	AOP	ARE	LPP	SSP	Category	Baseline	AOP	ARE	LPP	SSP
KungFu	99.3	**99.8**	99.8	99.8	99.8	Stamp	38.2	**70.5**	61.7	52.0	66.2
Cards	94.0	99.3	98.3	**99.8**	99.3	Bus	38.5	**58.2**	45.8	35.0	43.0
Dinosaur	87.8	98.7	98.5	97.2	**99.3**	Race Car	39.0	**60.3**	47.5	27.2	48.2
Fitness	85.8	95.8	90.8	**96.7**	91.0	Doll	37.5	**50.0**	45.5	32.0	32.0
Easter Egg	83.3	**95.7**	70.7	88.7	66.7	Old Car	35.3	**55.5**	49.8	36.5	47.3
PostCard	73.0	**94.8**	93.8	89.5	94.2	Tropical fish	40.0	**58.2**	46.5	40.5	42.7
Dish	67.5	**92.5**	79.5	90.0	87.5	Sunset	33.5	**56.5**	54.8	44.3	52.8
Horse	74.0	**95.0**	88.5	90.8	89.8	Leopard	51.0	**71.5**	48.7	44.0	45.8
owl	79.2	**80.3**	80.3	80.3	80.3	elephant	32.2	**49.5**	43.3	43.2	39.8
Flag	55.5	82.0	78.5	70.0	**82.0**	Surfing	32.5	**47.0**	45.8	25.5	45.8
Rodeo	53.3	**82.8**	59.7	58.5	67.5	Butterfly	28.5	**55.0**	50.3	33.5	46.3
Indoor decorate	51.7	**74.8**	53.0	45.7	48.0	Cat	24.8	45.0	**46.8**	31.3	38.8
Fruit	43.5	**74.8**	61.0	57.0	66.8	Lion	32.5	**52.5**	45.8	40.5	42.8
Cuisine	62.0	**78.8**	63.3	63.0	63.3	Bobsled	21.5	**42.3**	36.0	18.5	36.5
Antique	44.3	**66.8**	56.3	48.5	51.2	Ship	39.0	**57.0**	54.5	35.3	45.5
Tools	44.7	**75.3**	57.7	43.0	57.3	Bonsai	31.3	**49.0**	43.5	33.5	40.5
drink	36.3	**71.3**	68.0	40.5	69.5	Marble	26.3	49.3	**53.5**	40.5	44.0
Mosaic	44.3	**67.2**	61.0	57.5	60.5	Waves	31.5	**55.3**	52.3	22.0	48.8
Pyramid	43.0	**70.0**	67.3	58.0	63.5	Canvas	28.0	**46.8**	45.0	40.0	39.5
Firework	45.0	**77.8**	74.3	64.3	68.7	Tiger	38.5	**47.3**	35.5	26.0	24.8
flower	28.5	**67.0**	54.8	44.0	55.8	Orbit	23.8	**40.8**	36.0	28.0	39.5
Gun	46.2	**63.0**	51.0	47.7	52.0	Balloon	22.2	**36.5**	29.8	19.7	30.0
Mask	52.5	**72.3**	57.3	45.0	64.5	Train	25.5	**39.3**	33.0	15.5	30.2
Cell	38.5	**71.0**	62.5	47.2	62.2	Eagle	19.0	31.5	**34.3**	18.3	27.3
Mountain	53.3	**62.5**	45.3	39.5	48.5	Wolf	37.8	**38.2**	33.8	35.0	25.3

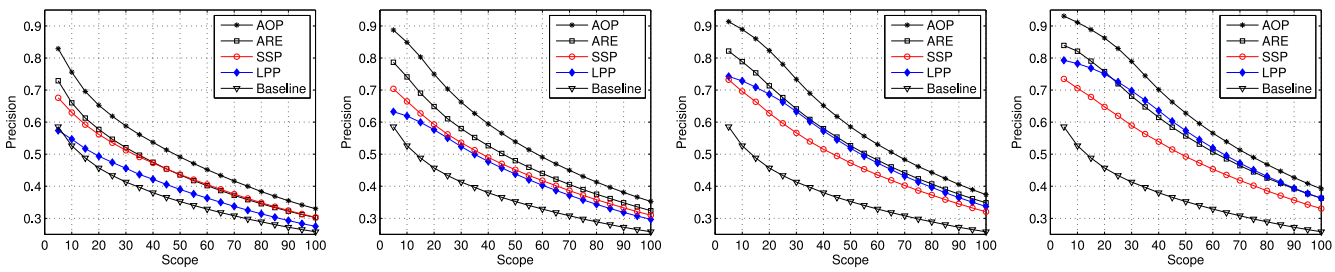
(relevant or irrelevant) is used for re-ranking. Since relevance feedback image retrieval is essentially a semi-supervised learning problem, we apply the LapRLS algorithm [15] to rank the images. Specifically, for a particular query, when the user feedback of "relevant" and "irrelevant" tags are given on some images, they are labeled as $+1$ and -1, respectively. Then the LapRLS algorithm is trained with those labels to distinguish between relevant and irrelevant images. In other words, the regression outputs are used to rank the candidate images. Note that the images that have been selected at previous iterations are excluded from later selections. For each query, the automatic relevance feedback mechanism is performed for four iterations.

In order to reduce the computational complexity, we do not take all the images in the database to compute the projection matrix. Instead, we only take the top 300 images at the previous retrieval iteration, plus the labeled images, to find the optimal projection. We empirically set the two parameters λ_1 and λ_2 to 10^{-4}. The same λ_1 and λ_2 are used in AOP and LapRLS. For dimensionality, we run the algorithms for all the 2 to 50 dimensions and report the best results.

5.4 Image Retrieval Performance

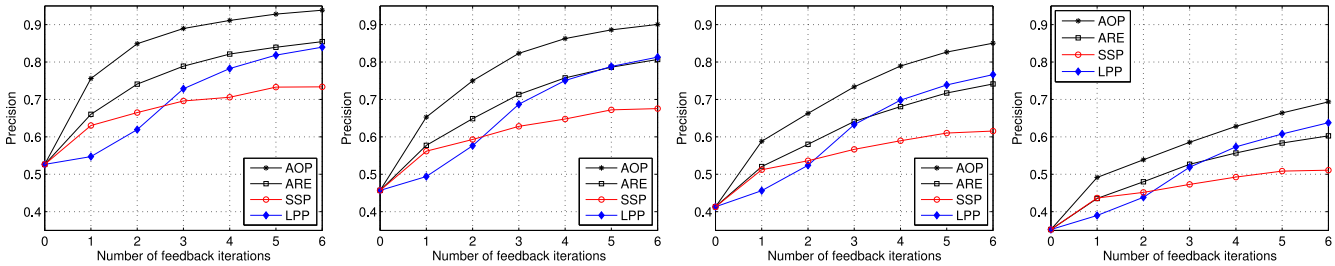
In the real world, it is not practical to require the user to provide many rounds of feedback. The retrieval performance after the first round of feedbacks is the most important. Table 1 shows the precision at the top 20 after the first round of feedbacks for all the 50 categories. The *baseline* method describes the initial retrieval result without feedback information. Specifically, at the beginning of retrieval, the Euclidean distances in the original 128-dimensional space are used to rank the images in the database. After the user provides relevance feedbacks, the AOP, ARE, SSP, and LPP algorithms are then applied to re-rank the images in the database. As can be seen, among all the 50 categories, AOP performs the best on 43 categories. The average retrieval precisions for AOP, ARE, SSP and LPP are 65.3, 57.7, 56.2, and 49.4 percent, respectively. Comparing to the second best algorithm, that is, ARE, our AOP algorithm achieves 7.6 percent improvement. Moreover, it would be important to note that, on all the 50 categories, our algorithm performs better than the baseline approach. However, for the other algorithms, on some categories, they may perform even worse than the baseline approach which does not use any relevance feedbacks. The reason is because AOP is explicitly designed for LapRLS, and thus when combined with LapRLS, it delivers the best result.

Fig. 1 shows the average *precision-scope* curves of the different algorithms for the first four feedback iterations. As can be seen, our AOP algorithms performs much better than the other three algorithms on the entire scope. After the first feedback iteration, LPP performs the worst. ARE performs slightly better than SSP, especially when the scope is small. After the second feedback iteration,



(a) Feedback Iteration 1 (b) Feedback Iteration 2 (c) Feedback Iteration 3 (d) Feedback Iteration 4

Fig. 1. The average *precision-scope* curves of different algorithms for the first four feedback iterations. The AOP algorithm performs the best on the entire scope.



(a) Precision at Top 10 (b) Precision at Top 20 (c) Precision at Top 30 (d) Precision at Top 50

Fig. 2. Performance evaluation of the five learning algorithms for relevance feedback image retrieval. (a) Precision at top 10, (b) Precision at top 20, and (c) Precision at top 30.

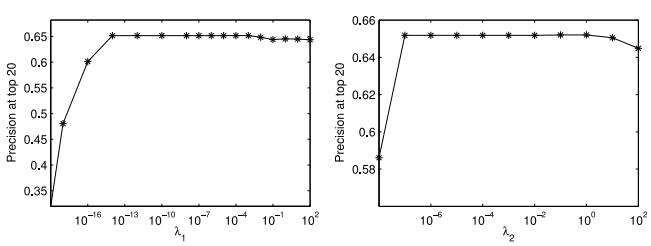
the performance of both AOP, ARE and LPP increases significantly. ARE consistently outperforms SSP and LPP is only slightly worse than SSP. After the third and forth iterations, both ARE and LPP outperform SSP. All of these four algorithms AOP, ARE, SSP, and LPP are significantly better than the baseline, which indicates that the user-provided relevance feedbacks are very helpful in improving the retrieval performance. By iteratively adding the user's feedbacks, the corresponding precisions (at the top 10, top 20, top 30, and top 50) of the four algorithms are, respectively, shown in Fig. 2. We can see that our AOP algorithm performs the best for all the cases. The performance improves significantly as the number of relevance feedbacks increases. For precision at top 20, the retrieval precision increases from 45.7 to 90.1 percent after six feedback iterations. For ARE and LPP, the retrieval precisions increase from 45.7 to 80.7 percent and 81.3 percent, respectively. For SSP, there is no convincing evidence that it can take full advantage of more relevance feedbacks.

5.5 Parameter Selection

In this section, we evaluate the sensitivity of our algorithm with respect to the two parameters λ_1 and λ_2. We run the same five-fold cross validation with different values of the parameters. The averaged precisions at top 20 are shown in Figs. 3a and 3b. Specifically, in Fig. 3a, we fix $\lambda_2 = 10^{-4}$ as in previous experiments and let λ_1 vary; In Fig. 3b, we fix $\lambda_1 = 10^{-4}$ as in previous experiments and let λ_2 vary. As we can see, our algorithm is quite stable with respect to the two parameters in a wide range.

6 Conclusions

This paper presents a novel linear dimensionality reduction algorithm, called A-Optimal Projection, from the perspective of statistical design. Unlike traditional linear dimensionality reduction algorithms which are not directly related to the classification task, our approach aims to minimize the prediction error of a regression model while reducing the dimensionality. In comparison with three state-of-the-art methods, the experimental results validate that the new method achieve significantly higher accuracy for image retrieval.



(a) λ_1 varies and λ_2 is fixed to 10^{-4} (b) λ_1 is fixed to 10^{-4} and λ_2 varies

Fig. 3. Precision at top 20 for AOP with different values of the two parameters.

Acknowledgments

This work was supported by National Basic Research Program of China (973 Program) under Grant 2012CB316400, National Program for Special Support of Top-Notch Young Professionals, and National Natural Science Foundation of China under Grant 61233011 and Grant 61125203.

References

[1] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[2] W. Liu, D. Tao, and J. Liu, "Transductive component analysis," in *Proc. IEEE Int. Conf. Data Min.*, Pisa, Italy, 2008, pp. 433–442.

[3] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[4] Y. Chen, J. Z. Wang, and R. Krovetz, "Content-based image retrieval by clustering," in *Proc. 5th ACM SIGMM Int. Workshop Multimedia Information Retrieval*, Berkeley, CA, USA, 2003, pp. 193–200.

[5] S. Gordon, H. Greenspan, and J. Goldberger, "Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nice, France, 2003, pp. 370–377.

[6] X. He, W.-Y. Ma, and H.-J. Zhang, "Learning an image manifold for retrieval," in *Proc. 12th Annu. ACM Conf. Multimedia*, New York, NY, USA, Oct. 2004, pp. 17–23.

[7] W. Liu, W. Jiang, and S.-F. Chang, "Relevance aggregation projections for image retrieval," in *Proc. ACM Int. Conf. Image Video Retrieval*, Niagara Falls, Canada, 2008, pp. 119–126.

[8] D. Tao, X. Tang, X. Li, and Y. Rui, "Kernel direct biased discriminant analysis: A new content-based image retrieval relevance feedback algorithm," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 716–727, Aug. 2006.

[9] L. Zhang, C. Chen, J. Bu, Z. Chen, S. Tan, and X. He, "Discriminative codeword selection for image representation," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 173–182.

[10] Y.-Y. Lin, T.-L. Liu, and H.-T. Chen, "Semantic manifold learning for image retrieval," in *Proc. 13th Annu. Int. ACM Conf. Multimedia*, Singapore, Nov. 2005, pp. 249–258.

[11] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[12] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[13] X. Li, S. Lin, S. Yan, and D. Xu, "Discriminant locally linear embedding with high-order tensor data," *IEEE Trans. Syst., Man, Cybern., Part B*, vol. 38, no. 2, pp. 342–352, Apr. 2008.

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Information Process. Syst. 14*, Cambridge, MA, USA, 2001, pp. 585–591.

[15] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[16] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*. London, U.K.: Oxford Univ. Press, 2007.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[18] J. Nilsson, F. Sha, and M. I. Jordan, "Regression on manifolds using kernel dimension reduction," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 697–704.

[19] M. Kim and V. Pavlovic, "Dimensionality reduction using covariance operator inverse regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[20] A. Aswani, P. Bickel, and C. Tomlin, "Regression on manifolds: Estimation of the exterior derivative," *Ann. Statist.*, vol. 39, no. 1, pp. 48–81, 2011.

[21] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst. 16*, Cambridge, MA, USA, 2003, pp. 153–160.
[22] J. Yu and Q. Tian, "Learning image manifold by semantic subspace projection," in *Proc. 14th Annu. Int. ACM Conf. Multimedia*, Santa Barbara, CA, USA, Oct., 2006, pp. 297–306.
[23] F. R. K. Chung, *Spectral Graph Theory* (Series Regional Conference Series in Mathematics), Providence, RI, USA: AMS, vol. 92, 1997.
[24] D. Cai, X. He, and J. Han, "Spectral regression for efficient regularized subspace learning," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[25] I. T. Jolliffe, "A note on the use of principal components in regression," *J. Roy. Statist. Soc.*, vol. 31, no. 3, pp. 300–303, 1982.
[26] P. Geladi and B. R. Kowalski. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta* [Online]. *185*, pp. 1–17. Available: http://www.sciencedirect.com/science/article/pii/0003267086800289
[27] K.-C. Li. (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* [Online]. *86(414)*, pp. 316–327. Available: http://www.jstor.org/stable/2290563
[28] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Kernel dimension reduction in regression," *Ann. Statist.*, vol. 37, no. 4, pp. 1871–1905, 2009.
[29] L. Zhang, C. Chen, W. Chen, J. Bu, D. Cai, and X. He, "Convex experimental design using manifold structure for image retrieval," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 45–53.
[30] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.
[31] H. Yu, M. Li, H.-J. Zhang, and J. Feng, "Color texture moments for content-based image retrieval," in *Proc. Int. Conf. Image Process.*, 2002, pp. 24–28.
[32] D. P. Huijsmans and N. Sebe, "How to complete performance graphs in content-based image retrieval: Add generality and normalize scope," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 245–251, Feb. 2005.