Laplacian Optimal Design for Image Retrieval

Xiaofei He
Yahoo!
hex@yahoo-inc.com

Wanli Min
IBM research
wanlimin@us.ibm.com

Deng Cai
CS Dept., UIUC
dengcai2@uiuc.edu

Kun Zhou
Microsoft Research Asia
kunzhou@microsoft.com

ABSTRACT

Relevance feedback is a powerful technique to enhance Content-Based Image Retrieval (CBIR) performance. It solicits the user's relevance judgments on the retrieved images returned by the CBIR systems. The user's labeling is then used to learn a classifier to distinguish between relevant and irrelevant images. However, the top returned images may not be the most informative ones. The challenge is thus to determine which unlabeled images would be the most informative (i.e., improve the classifier the most) if they were labeled and used as training samples. In this paper, we propose a novel active learning algorithm, called **Laplacian Optimal Design** (LOD), for relevance feedback image retrieval. Our algorithm is based on a regression model which minimizes the least square error on the measured (or, labeled) images and simultaneously preserves the local geometrical structure of the image space. Specifically, we assume that if two images are sufficiently close to each other, then their measurements (or, labels) are close as well. By constructing a nearest neighbor graph, the geometrical structure of the image space can be described by the graph Laplacian. We discuss how results from the field of optimal experimental design may be used to guide our selection of a subset of images, which gives us the most amount of information. Experimental results on Corel database suggest that the proposed approach achieves higher precision in relevance feedback image retrieval.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information search and retrieval—*Relevance feedback*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Experimental design*

General Terms

Algorithms, Performance, Theory

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
Copyright 2007 ACM 978-1-59593-597-7/07/0007 ...$5.00.

Keywords

Image retrieval, active learning, experimental design, relevance feedback, regression

1. INTRODUCTION

In many machine learning and information retrieval tasks, there is no shortage of unlabeled data but labels are expensive. The challenge is thus to determine which unlabeled samples would be the most informative (i.e., improve the classifier the most) if they were labeled and used as training samples. This problem is typically called *active learning* [7]. Here the task is to minimize an overall cost, which depends both on the classifier accuracy and the cost of data collection. Many real world applications can be casted into active learning framework. Particularly, we consider the problem of relevance feedback driven Content-Based Image Retrieval (CBIR) [19].

Content-Based Image Retrieval (CBIR) has attracted substantial interests in the last decade [3], [4], [5], [8], [11], [14], [17], [18], [19], [20]. It is motivated by the fast growth of digital image databases which, in turn, require efficient search schemes. Rather than describe an image using text, in these systems an image query is described using one or more example images. The low level visual features (color, texture, shape, etc.) are automatically extracted to represent the images. However, the low level features may not accurately characterize the high level semantic concepts. To narrow down the semantic gap, relevance feedback is introduced into CBIR [18].

In many of the current relevance feedback driven CBIR systems, the user is required to provide his/her relevance judgments on the top images returned by the system. The labeled images are then used to train a classifier to separate images that match the query concept from those that do not. However, in general the top returned images may not be the most informative ones. In the worst case, all the top images labeled by the user may be positive and thus the standard classification techniques can not be applied due to the lack of negative examples. Unlike the standard classification problems where the labeled samples are pregiven, in relevance feedback image retrieval the system can actively select the images to label. Thus active learning can be naturally introduced into image retrieval.

Despite many existing active learning techniques, Support Vector Machine (SVM) active learning [21] and regression based active learning [1] have received the most interests. Based on the observation that the closer to the SVM bound-

ary an image is, the less reliable its classification is, SVM active learning selects those unlabeled images closest to the boundary to solicit user feedback so as to achieve maximal refinement on the hyperplane between the two classes. The major disadvantage of SVM active learning is that the estimated boundary may not be accurate enough. Moreover, it may not be applied at the beginning of the retrieval when there is no labeled images. Some other SVM based active learning algorithms can be found in [11], [13].

In statistics, the problem of selecting samples to label is typically referred to as *experimental design*. The sample $\mathbf{x}$ is referred to as *experiment*, and its label y is referred to as *measurement*. The study of *optimal experimental design* (OED) [1] is concerned with the design of experiments that are expected to minimize variances of a parameterized model. The intent of optimal experimental design is usually to maximize confidence in a given model, minimize parameter variances for system identification, or minimize the model's output variance. Classical experimental design approaches include *A-Optimal Design*, *D-Optimal Design*, and *E-Optimal Design*. All of these approaches are based on a least squares regression model. Comparing to SVM based active learning algorithms, experimental design approaches are much more efficient in computation. However, this kind of approaches takes only measured (or, labeled) data into account in their objective function, while the unmeasured (or, unlabeled) data is ignored.

Benefit from recent progresses on optimal experimental design and semi-supervised learning, in this paper we propose a novel active learning algorithm for image retrieval, called **Laplacian Optimal Design** (LOD). Unlike traditional experimental design methods whose loss functions are only defined on the measured points, the loss function of our proposed LOD algorithm is defined on both measured and unmeasured points. Specifically, we introduce a locality preserving regularizer into the standard least-square-error based loss function. The new loss function aims to find a classifier which is *locally* as smooth as possible. In other words, if two points are sufficiently close to each other in the input space, then they are expected to share the same label. Once the loss function is defined, we can select the most informative data points which are presented to the user for labeling. It would be important to note that the most informative images may not be the top returned images.

The rest of the paper is organized as follows. In Section 2, we provide a brief description of the related work. Our proposed Laplacian Optimal Design algorithm is introduced in Section 3. In Section 4, we compare our algorithm with the state-or-the-art algorithms and present the experimental results on image retrieval. Finally, we provide some concluding remarks and suggestions for future work in Section 5.

2. RELATED WORK

Since our proposed algorithm is based on regression framework. The most related work is optimal experimental design [1], including *A-Optimal Design*, *D-Optimal Design*, and *E-Optimal Design*. In this Section, we give a brief description of these approaches.

2.1 The Active Learning Problem

The generic problem of active learning is the following. Given a set of points $\mathcal{A} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$ in $\mathbb{R}^d$, find a subset $\mathcal{B} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_k\} \subset \mathcal{A}$ which contains the most informative points. In other words, the points $\mathbf{z}_i (i = 1, \cdots, k)$ can improve the classifier the most if they are labeled and used as training points.

2.2 Optimal Experimental Design

We consider a linear regression model

$$y = \mathbf{w}^T \mathbf{x} + \epsilon \tag{1}$$

where y is the *observation*, $\mathbf{x}$ 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 $\mathbf{w}$. Thus, the maximum likelihood estimate for the weight vector, $\hat{\mathbf{w}}$, is that which minimizes the sum squared error

$$J_{sse}(\mathbf{w}) = \sum_{i=1}^{k} \left(\mathbf{w}^T \mathbf{z}_i - y_i\right)^2 \tag{2}$$

The estimate $\hat{\mathbf{w}}$ gives us an estimate of the output at a novel input: $\hat{y} = \hat{\mathbf{w}}^T \mathbf{x}$.

By Gauss-Markov theorem, we know that $\hat{\mathbf{w}} - \mathbf{w}$ has a zero mean and a covariance matrix given by $\sigma^2 H_{sse}^{-1}$, where H_{sse} is the Hessian of $J_{sse}(\mathbf{w})$

$$H_{sse} = \left(\frac{\partial^2 J_{sse}}{\partial \mathbf{w}^2}\right) = \left(\sum_{i=1}^{k} \mathbf{z}_i \mathbf{z}_i^T\right) = ZZ^T$$

where $Z = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_k)$.

The three most common scalar measures of the size of the parameter covariance matrix in optimal experimental design are:

- D-optimal design: determinant of H_{sse}.
- A-optimal design: trace of H_{sse}.
- E-optimal design: maximum eigenvalue of H_{sse}.

Clearly, out of these three approaches, A-optimal design is the most efficient one. Some recent work on experimental design can be found in [10], [23].

3. LAPLACIAN OPTIMAL DESIGN

Since the covariance matrix H_{sse} used in traditional approaches is only dependent on the *measured* samples, i.e. $\mathbf{z}_i$'s, these approaches fail to evaluate the expected errors on the *unmeasured* samples. In this Section, we introduce a novel active learning algorithm called *Laplacian Optimal Design* (LOD) which makes efficient use of both measured (labeled) and unmeasured (unlabeled) samples.

3.1 The Objective Function

In many machine learning problems, it is natural to assume that if two points $\mathbf{x}_i$, $\mathbf{x}_j$ are sufficiently close to each other, then their measurements $(f(\mathbf{x}_i), f(\mathbf{x}_j))$ are close as well. Let S be a similarity matrix. Thus, a new loss function which respects the geometrical structure of the data space can be defined as follows:

$$J_0(\mathbf{w}) = \sum_{i=1}^{k} \left(f(\mathbf{z}_i) - y_i\right)^2 + \frac{\lambda}{2} \sum_{i,j=1}^{m} \left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)^2 S_{ij} \tag{3}$$

where y_i is the measurement (or, label) of $\mathbf{z}_i$. Note that, the loss function (3) is essentially the same as the one used

in Laplacian Regularized Regression (LRR, [2]). However, LRR is a passive learning algorithm where the training data is given. In this paper, we are focused on how to select the most informative data for training. The loss function with our choice of symmetric weights S_{ij} ($S_{ij} = S_{ji}$) incurs a heavy penalty if neighboring pints $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far apart. Therefore, minimizing $J_0(\mathbf{w})$ is an attempt to ensure that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are *close* then $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ are close as well. There are many choices of the similarity matrix S. A simple definition is as follows:

$$S_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ is among the } p \text{ nearest neighbors of } \mathbf{x}_j, \\ & \text{or } \mathbf{x}_j \text{ is among the } p \text{ nearest neighbors of } \mathbf{x}_i; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Let D be a diagonal matrix, $D_{ii} = \sum_j S_{ij}$, and $L = D - S$. The matrix L is called *graph Laplacian* in spectral graph theory [6]. Let $\mathbf{y} = (y_1, \cdots, y_k)^T$ and $X = (\mathbf{x}_1, \cdots, \mathbf{x}_m)$. Following some simple algebraic steps, we see that:

$$\begin{aligned} & J_0(\mathbf{w}) \\ = & \sum_{i=1}^{k} \left(\mathbf{w}^T\mathbf{z}_i - y_i\right)^2 + \frac{\lambda}{2}\sum_{i,j=1}^{m}\left(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right)^2 S_{ij} \\ = & \left(\mathbf{y} - Z^T\mathbf{w}\right)^T\left(\mathbf{y} - Z^T\mathbf{w}\right) + \lambda\mathbf{w}^T\Big(\sum_{i=1}^{m} D_{ii}\mathbf{x}_i\mathbf{x}_i^T \\ & - \sum_{i,j=1}^{m} S_{ij}\mathbf{x}_i\mathbf{x}_j^T\Big)\mathbf{w} \\ = & \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T Z\mathbf{y} + \mathbf{w}^T ZZ^T\mathbf{w} \\ & + \lambda\mathbf{w}^T\left(XDX^T - XSX^T\right)\mathbf{w} \\ = & \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T Z\mathbf{y} + \mathbf{w}^T\left(ZZ^T + \lambda XLX^T\right)\mathbf{w} \end{aligned}$$

The Hessian of $J_0(\mathbf{w})$ can be computed as follows:

$$\begin{aligned} H_0 &= \left(\frac{\partial^2 J_0}{\partial\mathbf{w}^2}\right) \\ &= ZZ^T + \lambda XLX^T \end{aligned}$$

In some cases, the matrix $ZZ^T + \lambda XLX^T$ is singular (e.g. if $m < d$). Thus, there is no stable solution to the optimization problem Eqn. (3). A common way to deal with this *ill-posed* problem is to introduce a Tikhonov regularizer into our loss function:

$$\begin{aligned} & J(\mathbf{w}) \\ = & \sum_{i=1}^{k}\left(\mathbf{w}^T\mathbf{z}_i - y_i\right)^2 + \frac{\lambda_1}{2}\sum_{i,j=1}^{m}\left(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right)^2 S_{ij} \\ & + \lambda_2\|\mathbf{w}\|^2 \end{aligned} \tag{5}$$

The Hessian of the new loss function is given by:

$$\begin{aligned} H &= \left(\frac{\partial^2 J}{\partial\mathbf{w}^2}\right) \\ &= ZZ^T + \lambda_1 XLX^T + \lambda_2 I \\ &:= ZZ^T + \Lambda \end{aligned}$$

where I is an identity matrix and $\Lambda = \lambda_1 XLX^T + \lambda_2 I$. Clearly, H is of full rank. Requiring that the gradient of $J(\mathbf{w})$ with respect to $\mathbf{w}$ vanish gives the optimal estimate $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} = H^{-1}Z\mathbf{y}$$

The following proposition states the bias and variance properties of the estimator for the coefficient vector $\mathbf{w}$.

PROPOSITION 3.1. *$E(\hat{\mathbf{w}} - \mathbf{w}) = -H^{-1}\Lambda\mathbf{w}$, $Cov(\hat{\mathbf{w}}) = \sigma^2(H^{-1} - H^{-1}\Lambda H^{-1})$*

PROOF. Since $\mathbf{y} = Z^T\mathbf{w} + \epsilon$ and $E(\epsilon) = 0$, it follows that

$$\begin{aligned} & E(\hat{\mathbf{w}} - \mathbf{w}) \qquad (6) \\ = & H^{-1}ZZ^T\mathbf{w} - \mathbf{w} \\ = & H^{-1}(ZZ^T + \Lambda - \Lambda)\mathbf{w} - \mathbf{w} \\ = & (I - H^{-1}\Lambda)\mathbf{w} - \mathbf{w} \\ = & -H^{-1}\Lambda\mathbf{w} \qquad (7) \end{aligned}$$

Notice $\mathrm{Cov}(\mathbf{y}) = \sigma^2 I$, the covariance matrix of $\hat{\mathbf{w}}$ has the expression:

$$\begin{aligned} \mathrm{Cov}(\hat{\mathbf{w}}) &= H^{-1}Z\mathrm{Cov}(\mathbf{y})Z^T H^{-1} \\ &= \sigma^2 H^{-1}ZZ^T H^{-1} \\ &= \sigma^2 H^{-1}(H - \Lambda)H^{-1} \\ &= \sigma^2(H^{-1} - H^{-1}\Lambda H^{-1}) \qquad (8) \end{aligned}$$

□

Therefore mean squared error matrix for the coefficients $\mathbf{w}$ is

$$\begin{aligned} & E(\mathbf{w} - \hat{\mathbf{w}})(\mathbf{w} - \hat{\mathbf{w}})^T \qquad (9) \\ = & H^{-1}\Lambda\mathbf{w}\mathbf{w}^T\Lambda H^{-1} + \sigma^2(H^{-1} - H^{-1}\Lambda H^{-1}) \qquad (10) \end{aligned}$$

For any $\mathbf{x}$, let $\hat{y} = \hat{\mathbf{w}}^T\mathbf{x}$ be its predicted observation. The expected squared prediction error is

$$\begin{aligned} & E(y - \hat{y})^2 \\ = & E(\epsilon + \mathbf{w}^T\mathbf{x} - \hat{\mathbf{w}}^T\mathbf{x})^2 \\ = & \sigma^2 + \mathbf{x}^T[E(\mathbf{w} - \hat{\mathbf{w}})(\mathbf{w} - \hat{\mathbf{w}})^T]\mathbf{x} \\ = & \sigma^2 + \mathbf{x}^T[H^{-1}\Lambda\mathbf{w}\mathbf{w}^T\Lambda H^{-1} + \sigma^2 H^{-1} - \sigma^2 H^{-1}\Lambda H^{-1}]\mathbf{x} \end{aligned}$$

Clearly the expected square prediction error depends on the explanatory variable $\mathbf{x}$, therefore average expected square predictive error over the complete data set $\mathcal{A}$ is

$$\begin{aligned} & \frac{1}{m}\sum_{i=1}^{m} E(y_i - \hat{\mathbf{w}}^T\mathbf{x}_i)^2 \\ = & \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^T[H^{-1}\Lambda\mathbf{w}\mathbf{w}^T\Lambda H^{-1} + \sigma^2 H^{-1} - \sigma^2 H^{-1}\Lambda H^{-1}]\mathbf{x}_i \\ & + \sigma^2 \\ = & \frac{1}{m}\mathrm{Tr}(X^T[\sigma^2 H^{-1} + H^{-1}\Lambda\mathbf{w}\mathbf{w}^T\Lambda H^{-1} - \sigma^2 H^{-1}\Lambda H^{-1}]X) \\ & + \sigma^2 \end{aligned}$$

Since

$$\begin{aligned} & \mathrm{Tr}(X^T[H^{-1}\Lambda\mathbf{w}\mathbf{w}^T\Lambda H^{-1} - \sigma^2 H^{-1}\Lambda H^{-1}]X) \\ \ll & \mathrm{Tr}(\sigma^2 X^T H^{-1}X), \end{aligned}$$

Our Laplacian optimality criterion is thus formulated by minimizing the trace of $X^T H^{-1}X$.

Definition 1. Laplacian Optimal Design

$$\max_{Z=(\mathbf{z}_1,\cdots,\mathbf{z}_k)} \mathrm{Tr}\Big(X^T\left(ZZ^T + \lambda_1 XLX^T + \lambda_2 I\right)^{-1}X\Big) \tag{11}$$

where $\mathbf{z}_1, \cdots, \mathbf{z}_k$ are selected from $\{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$.

4. KERNEL LAPLACIAN OPTIMAL DESIGN

Canonical experimental design approaches (e.g. A-Optimal Design, D-Optimal Design, and E-Optimal) only consider linear functions. They fail to discover the intrinsic geometry in the data when the data space is highly nonlinear. In this section, we describe how to perform Laplacian Experimental Design in Reproducing Kernel Hilbert Space (RKHS) which gives rise to Kernel Laplacian Experimental Design (KLOD).

For given data points $\mathbf{x}_1, \cdots, \mathbf{x}_m \in \mathcal{X}$ with a positive definite mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a unique RKHS $\mathcal{H}_K$ of real valued functions on $\mathcal{X}$. Let $K_t(s)$ be the function of s obtained by fixing t and letting $K_t(s) \doteq K(s,t)$. $\mathcal{H}_K$ consists of all finite linear combinations of the form $\sum_{i=1}^{l} \alpha_i K_{t_i}$ with $t_i \in \mathcal{X}$ and limits of such functions as the t_i become dense in $\mathcal{X}$. We have $\langle K_s, K_t \rangle_{\mathcal{H}_K} = K(s,t)$.

4.1 Derivation of LOD in Reproducing Kernel Hilbert Space

Consider the optimization problem (5) in RKHS. Thus, we seek a function $f \in \mathcal{H}_K$ such that the following objective function is minimized:

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{k} \big(f(\mathbf{z}_i) - y_i\big)^2 + \frac{\lambda_1}{2} \sum_{i,j=1}^{m} \big(f(\mathbf{x}_i) - f(\mathbf{x}_j)\big)^2 S_{ij} + \lambda_2 \|f\|^2 \tag{12}$$

We have the following proposition.

PROPOSITION 4.1. *Let $\mathcal{H} = \{\sum_{i=1}^{m} \alpha_i K(\cdot, \mathbf{x}_i) | \alpha_i \in \mathbb{R}\}$ be a subspace of $\mathcal{H}_K$, the solution to the problem (12) is in $\mathcal{H}$.*

PROOF. Let $\mathcal{H}^{\perp}$ be the orthogonal complement of $\mathcal{H}$, i.e. $\mathcal{H}_K = \mathcal{H} \oplus \mathcal{H}^{\perp}$. Thus, for any function $f \in \mathcal{H}_K$, it has orthogonal decomposition as follows:

$$f = f_{\mathcal{H}} + f_{\mathcal{H}^{\perp}}$$

Now, let's evaluate f at $\mathbf{x}_i$:

$$\begin{aligned} f(\mathbf{x}_i) &= \langle f, K_{\mathbf{x}_i} \rangle_{\mathcal{H}_K} \\ &= \langle f_{\mathcal{H}} + f_{\mathcal{H}_{\perp}}, K_{\mathbf{x}_i} \rangle_{\mathcal{H}_K} \\ &= \langle f_{\mathcal{H}}, K_{\mathbf{x}_i} \rangle_{\mathcal{H}_K} + \langle f_{\mathcal{H}^{\perp}}, K_{\mathbf{x}_i} \rangle_{\mathcal{H}_K} \end{aligned}$$

Notice that $K_{\mathbf{x}_i} \in \mathcal{H}$ while $f_{\mathcal{H}^{\perp}} \in \mathcal{H}^{\perp}$. This implies that $\langle f_{\mathcal{H}^{\perp}}, K_{\mathbf{x}_i} \rangle_{\mathcal{H}_K} = 0$. Therefore,

$$f(\mathbf{x}_i) = \langle f_{\mathcal{H}}, K_{\mathbf{x}_i} \rangle_{\mathcal{H}_K} = f_{\mathcal{H}}(\mathbf{x}_i)$$

This completes the proof. □

Proposition 4.1 tells us the minimizer of problem (12) admits a representation $f^* = \sum_{i=1}^{m} \alpha_i K(\cdot, \mathbf{x}_i)$.

Let $\phi : \mathbb{R}^d \to \mathcal{H}$ be a feature map from the input space $\mathbb{R}^d$ to $\mathcal{H}$, and $K(\mathbf{x}_i, \mathbf{x}_j) = < \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) >$. Let $\mathbf{X}$ denote the data matrix in RKHS, $\mathbf{X} = (\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_m))$. Similarly, we define $\mathbf{Z} = (\phi(\mathbf{z}_1), \phi(\mathbf{z}_2), \cdots, \phi(\mathbf{z}_k))$. Thus, the optimization problem in RKHS can be written as follows:

$$\min_{\mathbf{Z}} tr\Big(\mathbf{X}^T \big(\mathbf{Z}\mathbf{Z}^T + \lambda_1 \mathbf{X} L \mathbf{X}^T + \lambda_2 I\big)^{-1} \mathbf{X}\Big) \tag{13}$$

Since the mapping function ϕ is generally unknown, there is no direct way to solve problem (13). In the following, we apply kernel tricks to solve this optimization problem. Let $\mathbf{X}^{-1}$ be the *Moore-Penrose* inverse (also known as *pseudo* inverse) of $\mathbf{X}$. Thus, we have:

$$\begin{aligned} & \mathbf{X}^T \big(\mathbf{Z}\mathbf{Z}^T + \lambda_1 \mathbf{X} L \mathbf{X}^T + \lambda_2 I\big)^{-1} \mathbf{X} \\ = \; & \mathbf{X}^T \mathbf{X} \mathbf{X}^{-1} \big(\mathbf{Z}\mathbf{Z}^T + \lambda_1 \mathbf{X} L \mathbf{X}^T + \lambda_2 I\big)^{-1} (\mathbf{X}^T)^{-1} \mathbf{X}^T \mathbf{X} \\ = \; & \mathbf{X}^T \mathbf{X} \big(\mathbf{Z}\mathbf{Z}^T \mathbf{X} + \lambda_1 \mathbf{X} L \mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{X}\big)^{-1} (\mathbf{X}^T)^{-1} \mathbf{X}^T \mathbf{X} \\ = \; & \mathbf{X}^T \mathbf{X} \big(\mathbf{X}^T \mathbf{Z}\mathbf{Z}^T \mathbf{X} + \lambda_1 \mathbf{X}^T \mathbf{X} L \mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{X}^T \mathbf{X}\big)^{-1} \mathbf{X}^T \mathbf{X} \\ = \; & \mathbf{K}_{\mathbf{XX}} \big(\mathbf{K}_{\mathbf{XZ}} \mathbf{K}_{\mathbf{ZX}} + \lambda_1 \mathbf{K}_{\mathbf{XX}} L \mathbf{K}_{\mathbf{XX}} + \lambda_2 \mathbf{K}_{\mathbf{XX}}\big)^{-1} \mathbf{K}_{\mathbf{XX}} \end{aligned}$$

where $\mathbf{K}_{\mathbf{XX}}$ is a $m \times m$ matrix ($\mathbf{K}_{\mathbf{XX},ij} = K(\mathbf{x}_i, \mathbf{x}_j)$), $\mathbf{K}_{\mathbf{XZ}}$ is a $m \times k$ matrix ($\mathbf{K}_{\mathbf{XZ},ij} = K(\mathbf{x}_i, \mathbf{z}_j)$), and $\mathbf{K}_{\mathbf{ZX}}$ is a $k \times m$ matrix ($\mathbf{K}_{\mathbf{ZX},ij} = K(\mathbf{z}_i, \mathbf{x}_j)$). Thus, the Kernel Laplacian Optimal Design can be defined as follows:

Definition 2. Kernel Laplacian Optimal Design

$$\min_{Z=(\mathbf{z}_1, \cdots, \mathbf{z}_k)} \; \mathrm{Tr}\Big(\mathbf{K}_{\mathbf{XX}} \big(\mathbf{K}_{\mathbf{XZ}} \mathbf{K}_{\mathbf{ZX}} + \lambda_1 \mathbf{K}_{\mathbf{XX}} L \mathbf{K}_{\mathbf{XX}} + \lambda_2 \mathbf{K}_{\mathbf{XX}}\big)^{-1} \mathbf{K}_{\mathbf{XX}}\Big) \tag{14}$$

4.2 Optimization Scheme

In this subsection, we discuss how to solve the optimization problems (11) and (14). Particularly, if we select a linear kernel for KLOD, then it reduces to LOD. Therefore, we will focus on problem (14) in the following.

It can be shown that the optimization problem (14) is NP-hard. In this subsection, we develop a simple sequential greedy approach to solve (14). Suppose n points have been selected, denoted by a matrix $\mathbf{Z}^n = (\mathbf{z}_1, \cdots, \mathbf{z}_n)$. The $(n+1)$-*th* point $\mathbf{z}_{n+1}$ can be selected by solving the following optimization problem:

$$\max_{\mathbf{Z}^{n+1} = (\mathbf{Z}^n, \mathbf{z}_{n+1})} \; Tr\Big(\mathbf{K}_{\mathbf{XX}} \big(\mathbf{K}_{\mathbf{XZ}^{n+1}} \mathbf{K}_{\mathbf{Z}^{n+1}\mathbf{X}} + \lambda_1 \mathbf{K}_{\mathbf{XX}} L \mathbf{K}_{\mathbf{XX}} + \lambda_2 \mathbf{K}_{\mathbf{XX}}\big)^{-1} \mathbf{K}_{\mathbf{XX}}\Big) \tag{15}$$

The kernel matrices $\mathbf{K}_{\mathbf{XZ}^{n+1}}$ and $\mathbf{K}_{\mathbf{Z}^{n+1}\mathbf{X}}$ can be rewritten as follows:

$$\mathbf{K}_{\mathbf{XZ}^{n+1}} = \big(\mathbf{K}_{\mathbf{XZ}^n}, \mathbf{K}_{\mathbf{Xz}_{n+1}}\big), \mathbf{K}_{\mathbf{Z}^{n+1}\mathbf{X}} = \begin{pmatrix} \mathbf{K}_{\mathbf{Z}^n\mathbf{X}} \\ \mathbf{K}_{\mathbf{z}_{n+1}\mathbf{X}} \end{pmatrix}$$

Thus, we have:

$$\mathbf{K}_{\mathbf{XZ}^{n+1}} \mathbf{K}_{\mathbf{Z}^{n+1}\mathbf{X}} = \mathbf{K}_{\mathbf{XZ}^n} \mathbf{K}_{\mathbf{Z}^n\mathbf{X}} + \mathbf{K}_{\mathbf{Xz}_{n+1}} \mathbf{K}_{\mathbf{z}_{n+1}\mathbf{X}}$$

We define:

$$A = \mathbf{K}_{\mathbf{XZ}^n} \mathbf{K}_{\mathbf{Z}^n\mathbf{X}} + \lambda_1 \mathbf{K}_{\mathbf{XX}} L \mathbf{K}_{\mathbf{XX}} + \lambda_2 \mathbf{K}_{\mathbf{XX}}$$

A is only dependent on $\mathbf{X}$ and $\mathbf{Z}^n$. Thus, the $(n+1)$-*th* point $\mathbf{z}_{n+1}$ is given by:

$$\mathbf{z}_{n+1} = \arg\min_{\mathbf{z}_{n+1}} \mathrm{Tr}\Big(\mathbf{K}_{\mathbf{XX}} \big(A + \mathbf{K}_{\mathbf{Xz}_{n+1}} \mathbf{K}_{\mathbf{z}_{n+1}\mathbf{X}}\big)^{-1} \mathbf{K}_{\mathbf{XX}}\Big) \tag{16}$$

Each time we select a new point $\mathbf{z}_{n+1}$, the matrix A is updated by:

$$A \leftarrow A + \mathbf{K}_{\mathbf{Xz}_{n+1}} \mathbf{K}_{\mathbf{z}_{n+1}\mathbf{X}}$$

If the kernel function is chosen as inner product $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$, then $\mathcal{H}_K$ is a linear functional space and the algorithm reduces to LOD.

5. CONTENT-BASED IMAGE RETRIEVAL USING LAPLACIAN OPTIMAL DESIGN

In this section, we describe how to apply Laplacian Optimal Design to CBIR. We begin with a brief description of image representation using low level visual features.

5.1 Low-Level Image Representation

Low-level image representation is a crucial problem in CBIR. General visual features includes color, texture, shape, etc. Color and texture features are the most extensively used visual features in CBIR. Compared with color and texture features, shape features are usually described after images have been segmented into regions or objects. Since robust and accurate image segmentation is difficult to achieve, the use of shape features for image retrieval has been limited to special applications where objects or regions are readily available.

In this work, We combine 64-dimensional color histogram and 64-dimensional Color Texture Moment (CTM, [22]) to represent the images. The color histogram is calculated using $4 \times 4 \times 4$ bins in HSV space. The Color Texture Moment is proposed by Yu et al. [22], which 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 to describe different aspects of co-occurrence relations of image pixels in each channel of the (SVcosH, SVsinH, V) color space. Then CTM calculates the first and second moment of these maps as a representation of the natural color image pixel distribution. Please see [22] for details.

5.2 Relevance Feedback Image Retrieval

Relevance feedback is one of the most important techniques to narrow down the gap between low level visual features and high level semantic concepts [18]. Traditionally, the user's relevance feedbacks are used to update the query vector or adjust the weighting of different dimensions. This process can be viewed as an on-line learning process in which the image retrieval system acts as a learner and the user acts as a teacher. They typical retrieval process is outlined as follows:

1. The user submits a query image example to the system. The system ranks the images in database according to some pre-defined distance metric and presents to the user the top ranked images.

2. The system selects some images from the database and request the user to label them as "relevant" or "irrelevant".

3. The system uses the user's provided information to re-rank the images in database and returns to the user the top images. Go to step 2 until the user is satisfied.

Our Laplacian Optimal Design algorithm is applied in the second step for selecting the most informative images. Once we get the labels for the images selected by LOD, we apply Laplacian Regularized Regression (LRR, [2]) to solve the optimization problem (3) and build the classifier. The classifier is then used to re-rank the images in database. Note that, in order to reduce the computational complexity, we do not use all the unlabeled images in the database but only those within top 500 returns of previous iteration.

6. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of our proposed algorithm on a large image database. To demonstrate the effectiveness of our proposed LOD algorithm, we compare it with Laplacian Regularized Regression (LRR, [2]), Support Vector Machine (SVM), Support Vector Machine Active Learning (SVM_{active}) [21], and A-Optimal Design (AOD). Both SVM_{active}, AOD, and LOD are active learning algorithms, while LRR and SVM are standard classification algorithms. SVM only makes use of the labeled images, while LRR is a semi-supervised learning algorithm which makes use of both labeled and unlabeled images. For SVM_{active}, AOD, and LOD, 10 training images are selected by the algorithms themselves at each iteration. While for LRR and SVM, we use the top 10 images as training data. It would be important to note that SVM_{active} is based on the ordinary SVM, LOD is based on LRR, and AOD is based on the ordinary regression. The parameters λ_1 and λ_2 in our LOD algorithm are empirically set to be 0.001 and 0.00001. For both LRR and LOD algorithms, we use the same graph structure (see Eqn. 4) and set the value of p (number of nearest neighbors) to be 5. We begin with a simple synthetic example to give some intuition about how LOD works.

6.1 Simple Synthetic Example

A simple synthetic example is given in Figure 1. The data set contains two circles. Eight points are selected by AOD and LOD. As can be seen, all the points selected by AOD are from the big circle, while LOD selects four points from the big circle and four from the small circle. The numbers beside the selected points denote their orders to be selected. Clearly, the points selected by our LOD algorithm can better represent the original data set. We didn't compare our algorithm with SVM_{active} because SVM_{active} can not be applied in this case due to the lack of the labeled points.

6.2 Image Retrieval Experimental Design

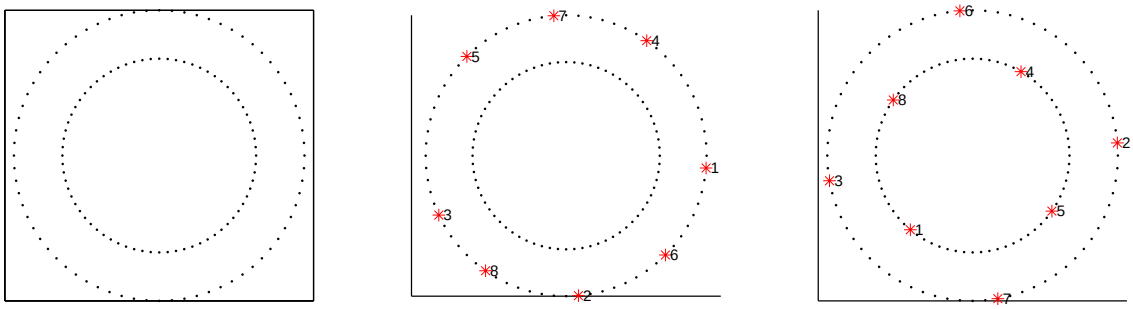
The image database we used consists of 7,900 images of 79 semantic categories, from COREL data set. It is a large and heterogeneous image set. Each image is represented as a 128-dimensional vector as described in Section 5.1. Figure 2 shows some sample images.

To exhibit the advantages of using our algorithm, we need a reliable way of evaluating the retrieval performance and the comparisons with other algorithms. We list different aspects of the experimental design below.

6.2.1 Evaluation Metrics

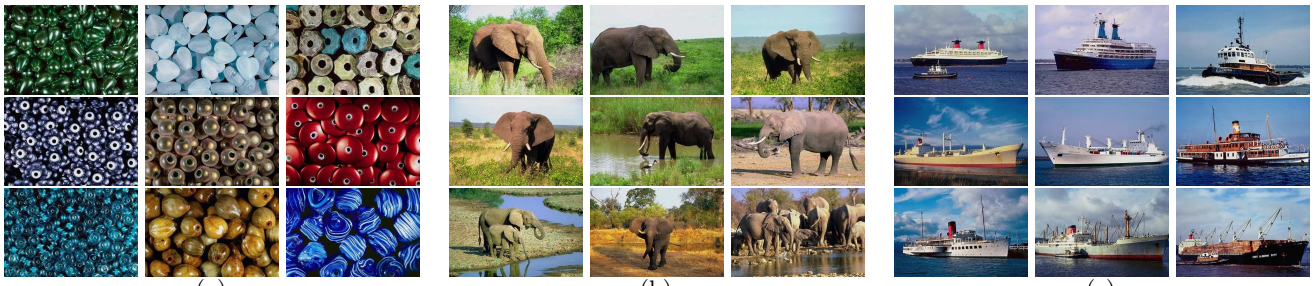
We use *precision-scope curve* and *precision rate* [15] to evaluate the effectiveness of the image retrieval algorithms. The scope is specified by the number (N) of top-ranked images presented to the user. The precision is the ratio of the number of relevant images presented to the user to the scope N. The precision-scope curve describes the precision with various scopes and thus gives an 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. Putting more images on a screen may affect the quality of the presented images. Therefore, the precision at top 20 ($N = 20$) is especially important.

In real world image retrieval systems, the query image is usually not in the image database. To simulate such environ-



(a) Data set (b) AOD (c) LOD

Figure 1: Data selection by active learning algorithms. The numbers beside the selected points denote their orders to be selected. Clearly, the points selected by our LOD algorithm can better represent the original data set. Note that, the SVM$_{active}$ algorithm can not be applied in this case due to the lack of labeled points.



(a) (b) (c)

Figure 2: Sample images from category *stone*, *elephant*, and *ship*.

ment, we use *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.

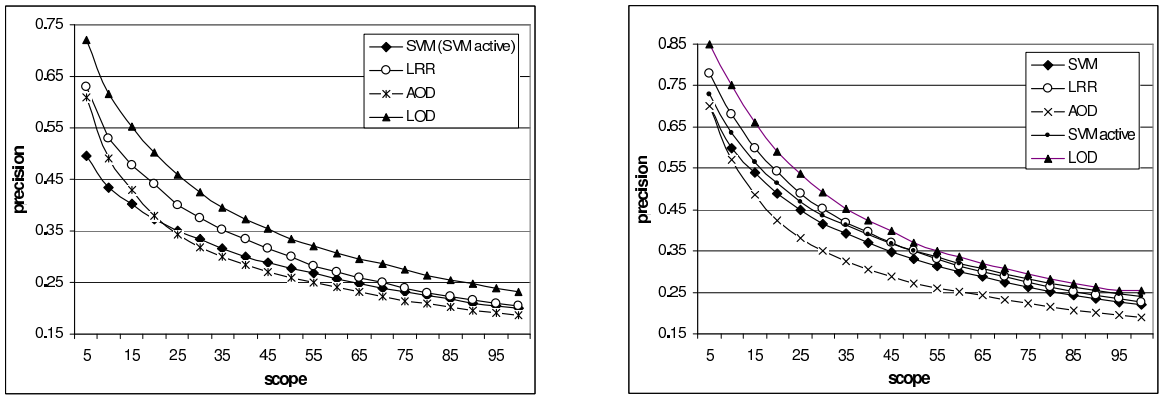
6.2.2 Automatic Relevance Feedback Scheme

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. 10 images were selected from the database for user labeling and the label information is used by the system for re-ranking. Note that, the images which have been selected at previous iterations are excluded from later selections. For each query, the automatic relevance feedback mechanism is performed for four iterations.

It is important to note that the automatic relevance feedback scheme used here is different from the ones described in [12], [16]. In [12], [16], the top four relevant and irrelevant images were selected as the feedback images. However, this may not be practical. In real world image retrieval systems, it is possible that most of the top-ranked images are relevant (or, irrelevant). Thus, it is difficult for the user to find both four relevant and irrelevant images. It is more reasonable for the users to provide feedback information only on the 10 images selected by the system.

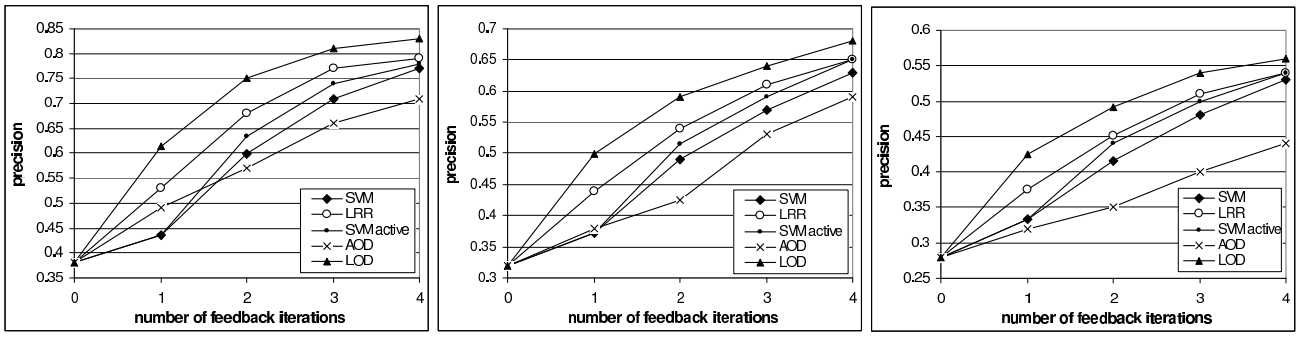
6.3 Image Retrieval Performance

In real world, it is not practical to require the user to provide many rounds of feedbacks. The retrieval performance after the first two rounds of feedbacks (especially the first round) is more important. Figure 3 shows the average *precision-scope* curves of the different algorithms for the first two feedback iterations. At the beginning of retrieval, the Euclidean distances in the original 128-dimensional space are used to rank the images in database. After the user provides relevance feedbacks, the LRR, SVM, SVM$_{active}$, AOD, and LOD algorithms are then applied to re-rank the images. In order to reduce the time complexity of active learning algorithms, we didn't select the most informative images from the whole database but from the top 500 images. For LRR and SVM, the user is required to label the top 10 images. For SVM$_{active}$, AOD, and LOD, the user is required to label 10 most informative images selected by these algorithms. Note that, SVM$_{active}$ can only be applied when the classifier is already built. Therefore, it can not be applied at the first round and we use the standard SVM to build the initial classifier. As can be seen, our LOD algorithm outperforms the other four algorithms on the entire scope. Also, the LRR algorithm performs better than SVM. This is because that the LRR algorithm makes efficient use of the unlabeled images by incorporating a locality preserving regularizer into the ordinary regression objective function. The AOD algorithm performs the worst. As the scope gets larger, the performance difference between these algorithms gets smaller.



(a) Feedback Iteration 1 (b) Feedback Iteration 2

Figure 3: The average *precision-scope* curves of different algorithms for the first two feedback iterations. The LOD algorithm performs the best on the entire scope. Note that, at the first round of feedback, the SVM$_{active}$ algorithm can not be applied. It applies the ordinary SVM to build the initial classifier.



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

Figure 4: 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. As can be seen, our LOD algorithm consistently outperforms the other four algorithms.

By iteratively adding the user's feedbacks, the corresponding precision results (at top 10, top 20, and top 30) of the five algorithms are respectively shown in Figure 4. As can be seen, our LOD algorithm performs the best in all the cases and the LRR algorithm performs the second best. Both of these two algorithms make use of the unlabeled images. This shows that the unlabeled images are helpful for discovering the intrinsic geometrical structure of the image space and therefore enhance the retrieval performance. In real world, the user may not be willing to provide too many relevance feedbacks. Therefore, the retrieval performance at the first two rounds are especially important. As can be seen, our LOD algorithm achieves 6.8% performance improvement for top 10 results, 5.2% for top 20 results, and 4.1% for top 30 results, comparing to the second best algorithm (LRR) after the first two rounds of relevance feedbacks.

6.4 Discussion

Several experiments on Corel database have been systematically performed. We would like to highlight several interesting points:

1. It is clear that the use of active learning is beneficial in the image retrieval domain. There is a significant increase in performance from using the active learning methods. Especially, out of the three active learning methods (SVM$_{active}$, AOD, LOD), our proposed LOD algorithm performs the best.

2. In many real world applications like relevance feedback image retrieval, there are generally two ways of reducing labor-intensive manual labeling task. One is active learning which selects the most informative samples to label, and the other is semi-supervised learning which makes use of the unlabeled samples to enhance the learning performance. Both of these two strategies have been studied extensively in the past [21], [11], [9], [12]. The work presented in this paper is focused on active learning, but it also takes advantage of the recent progresses on semi-supervised learning [2]. Specifically, we incorporate a locality preserving regularizer into the standard regression framework and find the most informative samples with respect to the new objective function. In this way, the active learning and semi-supervised learning techniques are seamlessly unified for learning an optimal classifier.

3. The relevance feedback technique is crucial to image

retrieval. For all the five algorithms, the retrieval performance improves with more feedbacks provided by the user.

7. CONCLUSIONS AND FUTURE WORK

This paper describes a novel active learning algorithm, called Laplacian Optimal Design, to enable more effective relevance feedback image retrieval. Our algorithm is based on an objective function which simultaneously minimizes the empirical error and preserves the local geometrical structure of the data space. Using techniques from experimental design, our algorithm finds the most informative images to label. These labeled images and the unlabeled images in the database are used to learn a classifier. The experimental results on Corel database show that both active learning and semi-supervised learning can significantly improve the retrieval performance.

In this paper, we consider the image retrieval problem on a small, static, and closed-domain image data. A much more challenging domain is the World Wide Web (WWW). For Web image search, it is possible to collect a large amount of user click information. This information can be naturally used to construct the affinity graph in our algorithm. Also, although our primary interest in this paper is focused on relevance feedback image retrieval, our results may also be of interest to researchers in patten recognition and machine learning, especially when a large amount of data is available but only a limited samples can be labeled.

8. REFERENCES

[1] A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs*. Oxford University Press, 2002.

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[3] J. Bi, Y. Chen, and J. Z. Wang. A sparse support vector machine approach to region based image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2004.

[4] G. Carneiro and N. Vasconcelos. A database centric view of semantic image and annotation and retrieval. In *Proc. International Conference on Research and Development in Information Retrieval (SIGIR'05)*, pages 559–566, Salvador, Brazil, 2005.

[5] E. Y. Chang, K. Goh, G. Sychay, and G. Wu. CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):26–38, January 2003.

[6] F. R. K. Chung. *Spectral Graph Theory*, volume 92 of *Regional Conference Series in Mathematics*. AMS, 1997.

[7] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[8] I. J. Cox, T. P. Minka, T. V. Papathomas, and P. N. Yianilos. The bayesian image retrieval system, pichunter: Theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, 9:20–37, 2000.

[9] A. Dong and B. Bhanu. A new semi-supervised em algorithm for image retrieval. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Madison, WI, 2003.

[10] P. Flaherty, M. I. Jordan, and A. P. Arkin. Robust design of biological experiments. In *Advances in Neural Information Processing Systems 18*, Vancouver, Canada, 2005.

[11] K.-S. Goh, E. Y. Chang, and W.-C. Lai. Multimodal concept-dependent active learning for image retrieval. In *Proceedings of the ACM Conference on Multimedia*, New York, October 2004.

[12] X. He. Incremental semi-supervised subspace learning for image retrieval. In *Proceedings of the ACM Conference on Multimedia*, New York, October 2004.

[13] S. C. Hoi and M. R. Lyu. A semi-supervised active learning framework for image retrieval. In *IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005.

[14] S. C. Hoi, M. R. Lyu, and R. Jin. A unified log-based relevance feedback scheme for image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):509–524, April 2006.

[15] D. P. Huijsmans and N. Sebe. How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):245–251, 2005.

[16] Y.-Y. Lin, T.-L. Liu, and H.-T. Chen. Semantic manifold learning for image retrieval. In *Proceedings of the ACM Conference on Multimedia*, Singapore, November 2005.

[17] W.-Y. Ma and B. S. Manjunath. Netra: a toolbox for navigating large image databases. *Multimedia Systems*, 7(3), May 1999.

[18] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interative content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5), 1998.

[19] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[20] J. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of the ACM Conference on Multimedia*, New York, 1996.

[21] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, 2001.

[22] H. Yu, M. Li, H.-J. Zhang, and J. Feng. Color texture moments for content-based image retrieval. In *International Conference on Image Processing*, pages 24–28, 2002.

[23] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *Proceedings of the 23^{rd} International Conference on Machine Learning*, Pittsburgh, PA, 2006.