Locally Regressive G-Optimal Design for Image Retrieval

Zheng-Jun Zha[†], Yan-Tao Zheng[‡], Meng Wang[†], Fei Chang[†], Tat-Seng Chua[†]
[†]School of Computing, National University of Singapore
[‡]Institute for Infocomm Research, Singapore
{zhazj, wangm, chuats}@comp.nus.edu.sg, yzheng@i2r.a-star.edu.sg

ABSTRACT

Content Based Image Retrieval (CBIR) has attracted increasing attention from both academia and industry. Relevance Feedback is one of the most effective techniques to bridge the semantic gap in CBIR. One of the key research problems related to relevance feedback is how to select the most informative images for users to label. In this paper, we propose a novel active learning algorithm, called Locally Regressive G-Optimal Design (LRGOD) for relevance feedback image retrieval. Our assumption is that for each image, its label can be well estimated based on its neighbors via a locally regressive function. LRGOD algorithm is developed based on a locally regressive least squares model which makes use of the labeled and unlabeled images, as well as simultaneously exploits the local structure of each image. The images that can minimize the maximum prediction variance are selected as the most informative ones. We evaluated the proposed LRGOD approach on two real-world image corpus: Corel and NUS-WIDE-OBJECT [5] datasets, and compare it to three state-of-the-art active learning methods. The experimental results demonstrate the effectiveness of the proposed approach.

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

Keywords

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

ICMR '11, April 17-20, Trento, Italy

1. INTRODUCTION

With the profusion of digital cameras, camera cell phones, as well as online photo sharing service (e.g., Flickr^{TM1}), large collection of images are now available to average users. But with that massive images within our reach, how do we go about finding the ones we want to see? Content-based image retrieval (CBIR) is a promising technique to address this issue of finding the images of interest by analyzing and comparing the content of all images in a database [6, 14]. It uses the low level visual feature to represent image content and retrieve images based on the similarity of their visual features. Most CBIR systems allow users to issue a query by means of uploading an example image, called query by example (QBE). Although CBIR has been extensively studied, one problem with it is the reliance on visual similarity for judging semantic similarity, which limits the retrieval performance due to the *semantic gap* between low-level visual features and high-level semantics [6, 14, 29, 7, 21].

Relevance feedback is the key technique to narrow down the semantic gap by exploiting users' interaction with CBIR systems [16]. In particular, users are encouraged to provide relevance judgements for the images retrieved by CBIR system, and relevance feedback algorithms are designed to learn and understand users' information needs from the feedbacks. In typical passive relevance feedback, the most relevant images (i.e., the top images returned by retrieval system) are presented for the users to give feedbacks. However, the feedbacks on these images do not provide the retrieval system with much additional information other than what it already knows about the users' information needs. Thus, one of the important research questions related to relevance feedback becomes how to select the most informative images to maximize the information gain.

Recently, active learning has been introduced into CBIR [13, 18]. By actively selecting a set of most informative images for user to label, the retrieval system is able to understand the user's information need and thus improve the retrieval performance effectively. The most popular active learning approach is Support Vector Machine active learning (SVM$_{active}$) [17]. SVM$_{active}$ builds a SVM classifier from feedback images to separate images that are relevant to user's query from those that are not. The informative samples are defined as the samples that are close to the SVM boundary, since they are usually the samples that the classifier is most uncertain about. One of the major shortcomings of SVM$_{active}$ is that the estimated boundary may not be accurate enough, especially when the labeled samples

[1]Flickr: http://www.flickr.com/

are insufficient, which is the case in relevance feedback.

The problem of selecting samples to label is typically referred to as *experimental design* in statistics. The sample $\mathbf{x}$ is referred to as *experiment*, and its label y is referred to as *measurement*. The study of Optimum Experiment Design (OED) [1] is concerned with the design of experiments that are expected to minimize variances of a parameterized model. There are two types of selection criteria for OED. One type is to select samples to minimize the variance of estimated model parameters, which results in A-,D- and E-Optimal Design. The other is to minimize the variance of the predication value, which results in I- and G-Optimal Design. All these algorihtms are based on a supervised learning algorithm: linear regression, which only takes into account the labeled data. In relevance feedback image retrieval, the labeled images are usually insufficient since users are not willing to provide too many feedbacks, while the unlabeled images are often plentiful. In this case, supervised learning algorithm may not perform well due to the lack of training data. Recent progress on semi-supervised learning shows the unlabeled data used with a small number of labeled samples can improve the learning performance greatly [30, 3, 20, 19, 26]. Motivated by this, new OED algorithms have been proposed based on semi-supervised learning framework [24, 12, 28, 11, 10, 4]. Based on Laplacian Regularized Least Squares(LapRLS) [2], Laplacian Regularized OED algorithms have been developed recently, including Laplacian Regularized A-, D-, I-, and G-Optimal Designs[12, 28, 11, 10, 4]. They select the images that can minimize the variance of the parameters or the prediction of LapRLS. All of these approaches are based on the assumption that if two samples are close to each other, their prediction values are close as well. Although the Laplacian Regularized OED performs more effectively than the classic OED, it is not able to make full use of the data distribution.

Motivated by the above observations, in this paper, we propose a novel Optimum Experimental Design algorithm, called Locally Regressive G-Optimal Design (LRGOD), for relevance feedback image retrieval. Our assumption is that for each sample, its label can be well estimated based on its neighbors via a locally regressive function. LRGOD is developed based on a locally regressive Least Squares model, which makes use of both the labeled and unlabeled data, as well as simultaneously exploits the local structure of each sample. The samples that can minimize the maximum prediction variance of the locally regressive Least Squares model are selected as the most informative samples and presented for the users to give feedbacks. We further perform LRGOD in the Reproducing Kernel Hilbert Space (RKHS) to obtain the nonlinear version. The experimental results on two real-world image corpus: Corel and NUS-WIDE-OBJECT datasets, demonstrate the effectiveness of the proposed approach.

The rest of this paper is organized as follows. In Section2, we provide a brief review of the related work. Our proposed Locally Regressive G-Optimal Design (LRGOD) is elaborated in Section 3. In Section 4, we describe how to perform LRGOD in Reproducing Kernel Hilbert Space. In Section 5, we compare the proposed LRGOD to the state-of-the-art active learning algorithms and report the experimental results on image retrieval. Finally, the conclusions and future works are discussed in Section 6.

2. RELATED WORK

As aforementioned, the most related work to the proposed approach is optimum experimental design. In this section, we will first provide a brief description of the generic active learning problem and then describe the conventional optimum experimental design.

2.1 The Active Learning Problem

Given a set of samples $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the d-dimensional feature vector, the generic problem of active learning is to find a subset $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K\} \subseteq \mathcal{X}$ containing the most informative samples $\mathbf{z}_k$. That is, if samples $\mathcal{Z}$ are labeled and used as training samples, the underlying classifier can predict the labels of the unlabeled samples most precisely [28].

2.2 Optimum Experimental Design

Optimum Experimental Design (OED) has received increasing attention due to its theoretical foundation and piratical effectiveness [1]. OED considers a linear regression model

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

where $\mathbf{w} \in \mathbb{R}^d$ are the model parameters, y is the real-valued output, and ε is the measurement noise which is a Gaussian random variable with zero mean and constant variance σ^2.

OED aims to choose the most informative samples $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K\}$ from $\mathcal{X}$ to learn a prediction function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ so that the expected prediction error can be minimized [1]. Given a set of selected samples $\mathcal{Z}$ and the corresponding labels $\mathcal{Y} = \{y_1, y_2, ..., y_K\}$, the linear regression model can be estimated following the *least squares* criteria:

$$\min_{\mathbf{w}} \sum_{i=1}^{K} (y_i - \mathbf{w}^T\mathbf{z}_i)^2. \tag{2}$$

The optimum solution is

$$\hat{\mathbf{w}} = (\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{y}, \tag{3}$$

where $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K]$ is the $d \times K$ feature matrix, and $\mathbf{y} = [y_1, y_2, ..., y_K]^T$ is the label vector. It can be proved [9] that $\hat{\mathbf{w}}$ is an unbiased estimation of $\mathbf{w}$ and its covariance matrix is given as

$$Cov(\hat{\mathbf{w}}) = \sigma^2(\mathbf{Z}\mathbf{Z}^T)^{-1}. \tag{4}$$

The conventional OED can be classified into two categories as according to the sample selection criteria. The first category is to select the samples that can minimize the size of the parameter covariance matrix, including D-, A-, and E-Optimal Design [1]. D-Optimal Design aims to select the samples to minimize the determinant of $Cov(\hat{\mathbf{w}})$ and thus minimize the volume of the confidence region. A-Optimal Design attempts to minimize the trace of $Cov(\hat{\mathbf{w}})$ and thus minimize the minimizes the dimensions of the enclosing box around the confidence region. E-Optimal Design minimizes the maximum eigenvalue of $Cov(\hat{\mathbf{w}})$, i.e., the size of the major axis of the confidence region. The other category aims to select the samples to minimize the variance of predication value, which results in I- and G-Optimal Design. For each sample $\mathbf{x}$, its prediction value is $\hat{\mathbf{w}}^T\mathbf{x}$ and the variance of the prediction value is $\mathbf{x}^T Cov(\hat{\mathbf{w}})\mathbf{x} = \sigma^2\mathbf{x}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{x}$. I-Optimal Design minimizes the average variance of prediction

value. G-Optimal Design chooses the most informative samples that can minimize the maximum variance of prediction value:

$$\underset{\mathcal{Z}\subseteq\mathcal{X}}{\operatorname{argmin}} \max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{x}. \quad (5)$$

As aforementioned, the conventional Optimum Experimental Design is based on supervised learning algorithm and usually suffers from insufficient training sample problem. To address this issue, Laplacian Regularized OED has been proposed to take into account both the labeled and unlabeled data, including Lapalcian Regularized A-, D-, I-, and G-Optimal Design [12, 10, 11, 28, 4]. All of these approaches are based on Laplacian Regularized Least Squares (LapRLS) [2], which assumes that if two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are close to each other, their prediction values $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ are close as well. They then select those samples that can minimize the variance of the parameters or the prediction variance of LapRLS. Laplacian Regularized OED mainly relies on an underlying k nearest neighborhoods graph, where the similarity among k nearest neighborhoods is set to 1, and 0 otherwise. Although it performs more effectively than the classic OED, it is not able to make full use of the data distribution.

3. LOCALLY REGRESSIVE G-OPTIMAL DESIGN

In this section, we will introduce a novel active learning algorithm, called Locally Regressive G-Optimal Design (LRGOD). Our approach is essentially based on the assumption that for each sample, its label can be well estimated based on its neighbors through a locally regressive function. The samples that can minimize the maximum predictive variance of a locally regressive Least Squares model are selected for labeling.

3.1 Locally Regressive Function

Recall that the linear regression model in Eq.(1), the predictive label of $\mathbf{x}_i$ is $f_i = \mathbf{w}^T\mathbf{x}_i$. Our assumption is that the value of f_i can be well estimated based on the neighbors of $\mathbf{x}_i$. Let $\mathcal{N}_k(\mathbf{x}_i) = \{\mathbf{x}_i, \mathbf{x}_{i_1}, \cdots, \mathbf{x}_{i_k}\}$ denote the sample $\mathbf{x}_i$ and its k-nearest neighbors. That is, f_i should approach to the output of a local model that is trained locally with the data $\{\mathbf{x}_j, f_j\}_{\mathbf{x}_j\in\mathcal{N}_k(\mathbf{x}_i)}$ [22, 23]. Here, we adopt a local linear regression model $o_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + b_i$, where $\mathbf{w}_i \in \mathbb{R}^d$ are the model parameters and $b_i \in \mathbb{R}$ is the bias. The local model $o_i(\mathbf{x})$ can be estimated by solving the following optimization problem.

$$\min_{\mathbf{w}_i, b_i} \sum_{\mathbf{x}_j\in\mathcal{N}_k(\mathbf{x}_i)} (\mathbf{w}_i^T\mathbf{x}_j + b_i - f_j)^2 + \eta \parallel \mathbf{w}_i \parallel^2, \quad (6)$$

where the regularization term $\parallel \mathbf{w}_i \parallel^2$ is used to avoid the "*overfitting.*"

By summing Eq. 6 over all the samples [23], we obtain

$$\min_{\{\mathbf{w}_i, b_i\}_{i=1}^N} \sum_{i=1}^{N} \{\parallel \mathbf{X}_i^T\mathbf{w}_i + b_i\mathbf{1}_{k+1} - \mathbf{f}_i \parallel^2 + \eta \parallel \mathbf{w}_i \parallel^2\}, \quad (7)$$

where $\mathbf{X}_i = [\mathbf{x}_i, \mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \cdots, \mathbf{x}_{i_k}]$ is the feature matrix containing all the samples in $\mathcal{N}_k(\mathbf{x}_i)$, $\mathbf{1}_{k+1} \in \mathbb{R}^{k+1}$ is a vector with all ones, and $\mathbf{f}_i = [f_i, f_{i_1}, f_{i_2}, \cdots, f_{i_k}]^T$.

By taking the derivative of the objective function with respect to $\mathbf{w}_i$ and b_i, we get the following solution:

$$\begin{aligned} \mathbf{w}_i &= (\mathbf{X}_i\mathbf{P}\mathbf{X}_i^T + \eta\mathbf{I})^{-1}\mathbf{X}_i\mathbf{P}\mathbf{f}_i \\ b_i &= \tfrac{1}{k+1}(\mathbf{1}_{k+1}^T\mathbf{f}_i - \mathbf{1}_{k+1}^T\mathbf{X}_i^T\mathbf{w}_i) \end{aligned} \quad (8)$$

where $\mathbf{P} = \mathbf{I} - \frac{1}{k+1}\mathbf{1}_{k+1}\mathbf{1}_{k+1}^T \in \mathbb{R}^{(k+1)\times(k+1)}$.

We substitute Eq.(8) for $\mathbf{w}_i$ and b_i and rewrite the objective function in Eq.(6) as

$$\begin{aligned} &\sum_{i=1}^{N} \{\parallel \mathbf{P}\mathbf{X}_i^T(\mathbf{X}_i\mathbf{P}\mathbf{X}_i^T + \eta\mathbf{I})^{-1}\mathbf{X}_i\mathbf{P}\mathbf{f}_i - \mathbf{P}\mathbf{f}_{(i)} \parallel^2 \\ &\quad + \eta\mathbf{f}_i^T\mathbf{P}\mathbf{X}_i^T(\mathbf{X}_i\mathbf{P}\mathbf{X}_i^T + \eta\mathbf{I})^{-2}\mathbf{X}_i\mathbf{P}\mathbf{f}_{(i)}\} \\ &= \sum_{i=1}^{N} \mathbf{f}_i^T(\mathbf{P} - \mathbf{P}\mathbf{X}_i^T(\mathbf{X}_i\mathbf{P}\mathbf{X}_i^T + \eta\mathbf{I})^{-1}\mathbf{X}_i\mathbf{P})\mathbf{f}_i \end{aligned} \quad (9)$$

Similar to [23], we introduce $\mathbf{O}_i \in \mathbb{R}^{N\times(k+1)}$ in which $\mathbf{O}_i(u, v) = 1$, if $u = \mathbf{q}_i(v)$, and $\mathbf{O}_i(u, v) = 0$, otherwise. $\mathbf{q}_i = [i, i_1, i_2, \cdots, i_k]^T$ are the sample indices in $\mathcal{N}_k(\mathbf{x}_i)$. Let $\mathbf{f} = [f_1, f_2, \cdots, f_N]^T$, we get $\mathbf{f}_i^T = \mathbf{f}^T\mathbf{O}_i$. Then, we can rewrite Eq.(9) as

$$\sum_{i=1}^{N} \mathbf{f}^T\mathbf{O}_i(\mathbf{P} - \mathbf{P}\mathbf{X}_i^T(\mathbf{X}_i\mathbf{P}\mathbf{X}_i^T + \eta\mathbf{I})^{-1}\mathbf{X}_i\mathbf{P})\mathbf{O}_i^T\mathbf{f} = \mathbf{f}^T\mathbf{L}\mathbf{f}. \quad (10)$$

where $\mathbf{L}$ is defined as

$$\mathbf{L} = [\mathbf{O}_1, \cdots, \mathbf{O}_N] \begin{bmatrix} \mathbf{L}_1 & & \\ & \cdots & \\ & & \mathbf{L}_N \end{bmatrix} [\mathbf{O}_1, \cdots, \mathbf{O}_N]^T; \quad (11)$$

$$\mathbf{L}_i = \mathbf{O}_i(\mathbf{P} - \mathbf{P}\mathbf{X}_i^T(\mathbf{X}_i\mathbf{P}\mathbf{X}_i^T + \eta\mathbf{I})^{-1}\mathbf{X}_i\mathbf{P})\mathbf{O}_i^T.$$

Recall that the linear regression model $f_i = \mathbf{w}^T\mathbf{x}_i$ in Eq.(1). Based on Eq.(10), we get the final locally regressive function as

$$\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w}. \quad (12)$$

3.2 Locally Regressive G-Optimal Design

As aforementioned, the classic OED and Laplacian regularised OED may not perform well due to the lack of training data or its limitation in making full use of data distribution. Motivated by this, we here propose a locally regressive least squares algorithm to take advantadge of both labeled and unlabeled data as well as simutaneously exploit the local structure of samples. Based on the least squares in Eq.(2) and the locally regressive function in Eq.(12), the locally regressive least squares algorithm is formulated as follows

$$\begin{aligned} J(\mathbf{w}) &= \sum_{i=1}^{K}(\mathbf{w}^T\mathbf{z}_i - y_i)^2 + \eta\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} + \gamma \parallel \mathbf{w} \parallel^2 \\ &= (\mathbf{Z}^T\mathbf{w} - \mathbf{y})^T(\mathbf{Z}^T\mathbf{w} - \mathbf{y}) + \eta\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} + \gamma \parallel \mathbf{w} \parallel^2, \end{aligned} \quad (13)$$

where $\mathbf{z}_i$ is the labeled sample in $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_1, \cdots, \mathbf{z}_K\}$ which is a subset of $\mathcal{X}$, $y_i \in \{1, -1\}$ is the label of $\mathbf{z}_i$. $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_K]$ is the $d \times K$ feature matrix, and $\mathbf{y} = [y_1, y_2, ..., y_K]^T$ is the label vector. The regularization term $\parallel \mathbf{w} \parallel^2$ is employed to control model complexity.

By setting the derivatives of $J(\mathbf{w})$ to zero with respect to $\mathbf{w}$, we get

$$(\mathbf{Z}\mathbf{Z}^T + \eta\mathbf{X}\mathbf{L}\mathbf{X}^T + \gamma\mathbf{I})\mathbf{w} - \mathbf{Z}\mathbf{y} = 0, \quad (14)$$

which leads to the optimal solution as

$$\hat{\mathbf{w}} = (\mathbf{Z}\mathbf{Z}^T + \eta\mathbf{X}\mathbf{L}\mathbf{X}^T + \gamma\mathbf{I})^{-1}\mathbf{Z}\mathbf{y} = \mathbf{H}_e^{-1}\mathbf{Z}\mathbf{y}, \quad (15)$$

where $\mathbf{H}_e = \frac{\partial^2 J}{\partial \mathbf{w}^2} = \mathbf{Z}\mathbf{Z}^T + \eta\mathbf{X}\mathbf{L}\mathbf{X}^T + \gamma\mathbf{I}$ is the hessian matrix of $J(\mathbf{w})$.

Given a test sample $\mathbf{x}$, the prediction value is $\hat{\mathbf{w}}^T\mathbf{x}$ with variance as:

$$\begin{aligned}\mathbf{x}^T Cov(\hat{\mathbf{w}})\mathbf{x} &= \mathbf{x}^T\mathbf{H}_e^{-1}\mathbf{Z}Cov(\mathbf{y})\mathbf{Z}^T\mathbf{H}_e^{-1}\mathbf{x} \\ &= \sigma^2\mathbf{x}^T\mathbf{H}_e^{-1}\mathbf{Z}\mathbf{Z}^T\mathbf{H}_e^{-1}\mathbf{x}.\end{aligned} \tag{16}$$

Our Locally Regressive G-Optimal Design (LRGOD) then selects the most informative samples to minimize the maximum variance of the prediction value.

$$\operatorname*{argmin}_{\mathcal{Z}\subseteq\mathcal{X}} \max_{\mathbf{x}_i}\{\mathbf{x}_i^T\mathbf{H}_e^{-1}\mathbf{Z}\mathbf{Z}^T\mathbf{H}_e^{-1}\mathbf{x}_i\}. \tag{17}$$

3.3 Solution

We adopt a sequential optimization approach to solve the Eq.(17) efficiently. Suppose a set of k samples $\mathcal{Z}_k = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_k\} \in \mathcal{X}$ have been selected as the k most informative samples. Let $\mathbf{Z}_k = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_k]$ denote the feature matrix. We define

$$\begin{aligned}\mathbf{H}_k &= \mathbf{Z}_k\mathbf{Z}_k^T + \eta\mathbf{X}\mathbf{L}\mathbf{X}^T + \gamma\mathbf{I} \\ \mathbf{H}_0 &= \eta\mathbf{X}\mathbf{L}\mathbf{X}^T + \gamma\mathbf{I}.\end{aligned} \tag{18}$$

The $(k+1)$-th sample $\mathbf{z}_{k+1}$ is selected as:

$$\begin{aligned}\mathbf{z}_{k+1} = \operatorname*{argmin}_{\mathbf{z}\in\mathcal{X}-\mathcal{Z}_k} \max_{\mathbf{x}_i}\{&\mathbf{x}_i^T(\mathbf{H}_k + \mathbf{z}\mathbf{z}^T)^{-1}(\mathbf{Z}_k\mathbf{Z}_k^T + \mathbf{z}\mathbf{z}^T) \\ &(\mathbf{H}_k + \mathbf{z}\mathbf{z}^T)^{-1}\mathbf{x}_i\}\end{aligned} \tag{19}$$

As can be seen, the most expensive calculation in Eq.(19) is the matrix inverse $(\mathbf{H}_k + \mathbf{z}\mathbf{z}^T)^{-1}$. Here we use the Sherman-Morrison-Woodbury formula [8] to avoid directly inverting a matrix. Given an invertible matrix $\mathbf{M}$, two column vectors $\mathbf{u}$ and $\mathbf{v}$, the Sherman-Morrison-Woodbury formula states:

$$(\mathbf{M} + \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{M}^{-1}}{1 + \mathbf{v}^T\mathbf{M}^{-1}\mathbf{u}}. \tag{20}$$

Thus, we can calculate $(\mathbf{H}_k + \mathbf{z}\mathbf{z}^T)^{-1}$ as

$$(\mathbf{H}_k + \mathbf{z}\mathbf{z}^T)^{-1} = \mathbf{H}_k^{-1} - \frac{\mathbf{H}_k^{-1}\mathbf{z}\mathbf{z}^T\mathbf{H}_k^{-1}}{1 + \mathbf{z}^T\mathbf{H}_k^{-1}\mathbf{z}}. \tag{21}$$

We thus only need to compute the inverse of $\mathbf{H}_0$. The inverse of $\mathbf{H}_{k+1}$ can be computed according to the above equation once $\mathbf{z}_{k+1}$ is found.

4. KERNEL LOCALLY REGRESSIVE G-OPTIMAL EXPERIMENT DESIGN

In this section, we develop a nonlinear version of LRGOD by performing experimental design in the Reproducing Kernel Hilbert Space (RKHS).

Let $K(\cdot,\cdot)$ denote a positive definite mercer kernel $\mathcal{K} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, and $\mathcal{H}_K$ be the corresponding Reproducing Kernel Hilbert Space (RKHS). Consider the optimization problem Eq.(13) in RKHS. We aim to find a function $f \in \mathcal{H}_\mathcal{K}$ to minimize the following objective function:

$$J(f) = \sum_{i=1}^{K}(f(\mathbf{z}_i) - y_i)^2 + \eta\mathbf{f}\mathbf{L}\mathbf{f}^T + \gamma \parallel f \parallel_\mathcal{H}^2, \tag{22}$$

where $\mathbf{f}$ is the vector $[f(\mathbf{x}_1), f(\mathbf{x}_2), \cdots, f(\mathbf{x}_N)]^T$, $\mathbf{f}\mathbf{L}\mathbf{f}^T$ is the locally regressive function, and $\parallel f \parallel_\mathcal{H}^2$ is a regularization term.

According to Representer Theorem [2], the optimal solution $\hat{f}$ is an expansion of kernel functions over both the labeled and unlabeled data:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{N}\hat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i). \tag{23}$$

Substituting this form into Eq.(22), we can arrive at the convex differentiable objective function of the N-dimensional variable $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_N]^T$:

$$\begin{aligned}\operatorname*{argmin}_{\boldsymbol{\alpha}}\{&(y - \mathbf{K}_{ZX}\boldsymbol{\alpha})^T(y - \mathbf{K}_{ZX}\boldsymbol{\alpha}) \\ &+ \eta\boldsymbol{\alpha}^T\mathbf{K}_{XX}\mathbf{L}\mathbf{K}_{XX}\boldsymbol{\alpha} + \gamma\boldsymbol{\alpha}^T\mathbf{K}_{XX}\boldsymbol{\alpha}\},\end{aligned} \tag{24}$$

where $\mathbf{K}_{XZ}$ is a $N \times K$ kernel matrix with $\mathbf{K}_{XZ}(i,j) = K(\mathbf{x}_i, \mathbf{z}_j)$, $\mathbf{K}_{XX}$ is a $N \times N$ kernel matrix with $\mathbf{K}_{XX}(i,j) = K(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathbf{K}_{ZX} = \mathbf{K}_{XZ}^T$.

By setting the derivatives of the objective function to zero with respect to $\boldsymbol{\alpha}$, we get the optimal solution as

$$\hat{\boldsymbol{\alpha}} = (\mathbf{K}_{XZ}\mathbf{K}_{ZX} + \eta\mathbf{K}_{XX}\mathbf{L}\mathbf{K}_{XX} + \gamma\mathbf{K}_{XX})^{-1}\mathbf{K}_{XZ}\mathbf{y}, \tag{25}$$

Denote $\mathbf{K}_{XZ}\mathbf{K}_{ZX} + \eta\mathbf{K}_{XX}\mathbf{L}\mathbf{K}_{XX} + \gamma\mathbf{K}_{XX}$ by $\mathbf{M}$. The covariance of $\hat{\boldsymbol{\alpha}}$ is given as

$$\begin{aligned}Cov(\hat{\boldsymbol{\alpha}}) &= \mathbf{M}^{-1}\mathbf{K}_{XZ}Cov(\mathbf{y})\mathbf{K}_{ZX}\mathbf{M}^{-1} \\ &= \sigma^2\mathbf{M}^{-1}\mathbf{K}_{XZ}\mathbf{K}_{ZX}\mathbf{M}^{-1}.\end{aligned} \tag{26}$$

Given a test sample $\mathbf{x}_i$, the prediction value is $\sum_{j=1}^{N}\hat{\alpha}_j K(\mathbf{x}_i, \mathbf{x}_j)$ with variance as:

$$\mathbf{k}_i^T Cov(\hat{\boldsymbol{\alpha}})\mathbf{k}_i = \sigma^2\mathbf{k}_i^T\mathbf{M}^{-1}\mathbf{K}_{XZ}\mathbf{K}_{ZX}\mathbf{M}^{-1}\mathbf{k}_i, \tag{27}$$

where $\mathbf{k}_i = [K(\mathbf{x}_i, \mathbf{x}_1), K(\mathbf{x}_i, \mathbf{x}_2), \cdots, K(\mathbf{x}_i, \mathbf{x}_N)]^T$ is the i-th column vector of $\mathbf{K}_{XX}$ corresponding to $\mathbf{x}_i$.

Thus, our kernel Locally Regressive G-Optimal Design is defined as

$$\operatorname*{argmin}_{\mathcal{Z}\subseteq\mathcal{X}} \max_{\mathbf{x}_i}\{\mathbf{k}_i^T\mathbf{M}^{-1}\mathbf{K}_{XZ}\mathbf{K}_{ZX}\mathbf{M}^{-1}\mathbf{k}_i\}. \tag{28}$$

This optimization problem can be solved through the following sequential optimization approach. Suppose k samples have been selected, we define

$$\begin{aligned}\mathbf{M}_k &= \mathbf{K}_{XZ_k}\mathbf{K}_{Z_kX} + \eta\mathbf{K}_{XX}\mathbf{L}\mathbf{K}_{XX} + \gamma\mathbf{K}_{XX} \\ \mathbf{M}_0 &= \eta\mathbf{K}_{XX}\mathbf{L}\mathbf{K}_{XX} + \gamma\mathbf{K}_{XX}.\end{aligned} \tag{29}$$

The $(k+1)$-th sample $\mathbf{z}_{k+1}$ can be found by solving the following problem:

$$\begin{aligned}\mathbf{z}_{k+1} = \operatorname*{argmin}_{\mathbf{z}\in\mathcal{X}-\mathcal{Z}_k} \max_{\mathbf{x}_i}\{&\mathbf{k}_i^T((\mathbf{M}_k + \mathbf{k}_z\mathbf{k}_z^T)^{-1} \\ &(\mathbf{K}_{XZ_k}\mathbf{K}_{Z_kX} + \mathbf{k}_z\mathbf{k}_z^T)(\mathbf{M}_k + \mathbf{k}_z\mathbf{k}_z^T)^{-1})\mathbf{k}_i\},\end{aligned} \tag{30}$$

where $\mathbf{k}_z = [K(\mathbf{z}, \mathbf{x}_1), K(\mathbf{z}, \mathbf{x}_2), \cdots, K(\mathbf{z}, \mathbf{x}_N)]^T$ is the corresponding column vector in $\mathbf{K}_{XX}$ of $\mathbf{z}$

By using Sherman-Morrison-Woodbury formula, the inverse of $\mathbf{M}_k + \mathbf{k}_z\mathbf{k}_z^T$ can be computed as

$$(\mathbf{M}_k + \mathbf{k}_z\mathbf{k}_z^T)^{-1} = \mathbf{M}_k^{-1} - \frac{\mathbf{M}_k^{-1}\mathbf{k}_z\mathbf{k}_z^T\mathbf{M}_k^{-1}}{1 + \mathbf{k}_z^T\mathbf{M}_k^{-1}\mathbf{k}_z} \tag{31}$$

5. EXPERIMENTS

We conducted experiments to evaluate the proposed approach and compare it to three state-of-the-art methods: Support Vector Machine active learning (SVM*active*) [17], G-Optimal Design (GOD) [1] and the recently proposed Laplacian Regularized G-Optimal Design (LapGOD) [4]. The performance comparison was conducted on two real-world image corpus: Corel and NUS-WIDE-OBJECT datasets, in the task of relevance feedback image retrieval.

5.1 Data and Methodology

The Corel image dataset we used consists of 5,000 images with 50 semantic categories, while the NUS-WIDE-OBJECT dataset consists of 30,000 images of 31 categories. Figure 1 shows some sample images from NUS-WIDE-OBJECT dataset. For the evaluation on Corel dataset, we randomly selected 20 images from each category as the query images. This gives rise to 1,000 query images in total. For the evaluation on NUS-WIDE OBJECT dataset, we randomly selected around 50 query images from each category and obtained 1,550 queries in total. To represent the image content, we extracted two types of visual features from each image: 225-dimensional block-wise color moment extracted over 5-by-5 division of the image, where each division was represented by a 9-dimensional feature, and 75-dimensional edge distribution histogram [15, 25]. We concatenated these features into a 330-dimensional feature vector.

For performance comparison, we applied the above four approaches (i.e., SVM*active*, GOD, LapGOD, and the proposed LRGOD) on relevance feedback image retrieval. The typical relevance feedback process is outlined as follows.

- Initially, the user submits a query image to the CBIR system. The system ranks the images in database according to certain pre-defined distance metric and presents to the user the top ranked images.
- The system then selects some images from the database and requests the user to label them as "relevant" or "irrelevant."
- The system uses the user's feedbacks to re-rank the images in the database then returns to the user the top images. Go to the second step until the user is satisfied.

Active learning approaches were applied in the second step to select the most informative images to label. For example, we used the proposed LRGOD algorithm to select the most informative images to label. After we obtained the labels of the selected images ("1" for relevant images and "-1" for irrelevant ones), a nonlinear model $f(\mathbf{x}) = \sum_{i=1}^{N} \hat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i)$ was estimated by minimizing the objective function in Eq.(22). The output of $f(\mathbf{x})$ is the relevance score of $\mathbf{x}$ with respect to the query. Then, all the images in database were re-ranked according to their relevance scores. We adopted RBF Kernel $K(\mathbf{x}_i, \mathbf{x}_j) = exp\{-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2\}$ in our implementation.

To reduce the computation cost, we selected the most informative images from the top 300 images returned by the retrieval system. In each feedback iteration, each of the four active learning approach (i.e.,SVM*active*, GOD, LapGOD, LRGOD) selected 10 images to label. Images that have been selected at previous feedback iterations were excluded from later selections. For each query, the relevance feedback was performed for five iterations. Note that SVM*active* cannot be applied directly at the first feedback iteration since there is no training samples. Alternatively, at the first iteration, we used the top 20 images as the training samples for SVM*active* and used the learned SVM classifier to re-rank the retrieval results. In our implementation, we did not use all the unlabeled images in the database but only those within the top 300 returns of previous iteration. The tradeoff parameters η and γ were set empirically and the number of neighbors was set to 5. The performance metrics *precision-scope* and *precision rate* were used to evaluate the retrieval performance [27]. 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 the relevant images within scope N. The precision-scope describes the precision with various scopes and thus gives an overall retrieval performance evaluation. On the other hand, the precision rate describes the precision at a particular value of scope. In general, users do not want to browse the retrieval results page by page and are only willing to browse the top-ranked results. Therefore, the precisions at top results are highly important, especially the precisions at top 10 and 20 results.

5.2 Experimental Results on Corel Dataset

Figure 2 illustrates the precision-scope of our proposed LRGOD algorithm and the three representative methods: SVM*active* [17], GOD [1], LapGOD [4] on the Corel dataset. Due to the space limitation, we only show the precision-scope for the first four feedback iterations in Fig. 2. By iteratively performing relevance feedback, the corresponding precision rates at top 10, 20, and 50 results of the four approaches are presented in Figure 3. From the experimental results, the following observations can be obtained:

- The proposed LRGOD algorithm outperforms the other three approaches on the entire scope and in every iteration. It especially performs well at the first feedback iteration. Since users are not willing to provide many feedbacks, the retrieval performance at the first feedback iteration is especially important. Consider the precision at top 10 results, LRGOD achieves 21.7%, 19.6%, and 24.3% improvements at the first iteration as compared to SVM*active*, GOD, and LapGOD, respectively.
- Both LRGOD and LapGOD are based on semi-supervised learning algorithm. By taking into account both the labeled and unlabeled data, LRGOD and LapGOD perform better than SVM*active* and GOD algorithms which only utilize the labeled samples and usually suffer from insufficient labeled sample problem.
- By exploiting the local structure of each sample through a locally regressive function, our LRGOD algorithm can outperform LapGOD method in every iteration. Specifically, LRGOD achieves 24.3%, 6.12%, 6.05%, 6.11%, and 6.07% improvements for top 10 results in the five iterations, respectively.
- The relevance feedback technique is beneficial to image retrieval. For all the four algorithms, the retrieval performance improves with more feedbacks provided by the user.

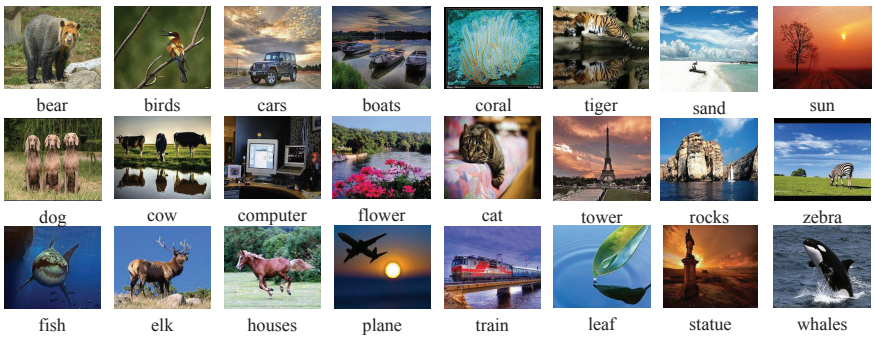

Figure 1: Sample images from NUS-WIDE-OBJECT dataset.

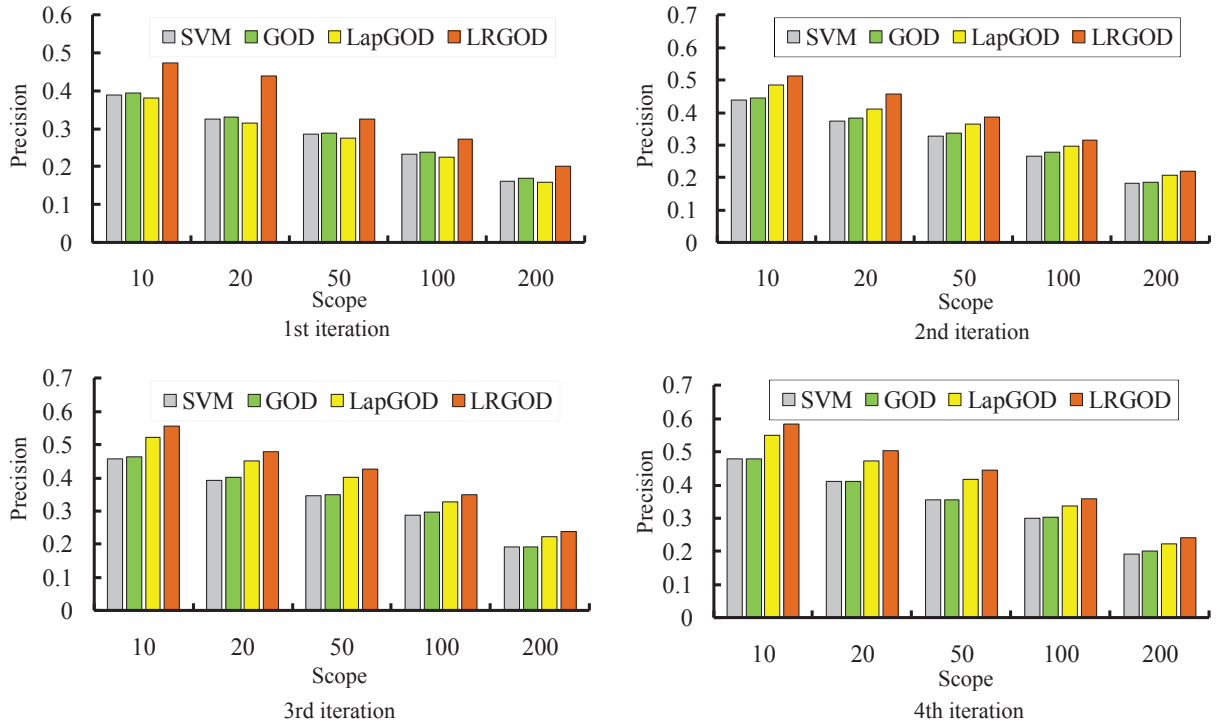

Figure 2: The performance of the four active learning approaches on Corel dataset measured by Precision-Scope. The four approaches are SVM*active* [17], G-Optimal Design (GOD) [1], Laplacian Regularized G-Optimal Design(LapGOD) [4], and the proposed Locally Regressive G-Optimal Design (LRGOD).

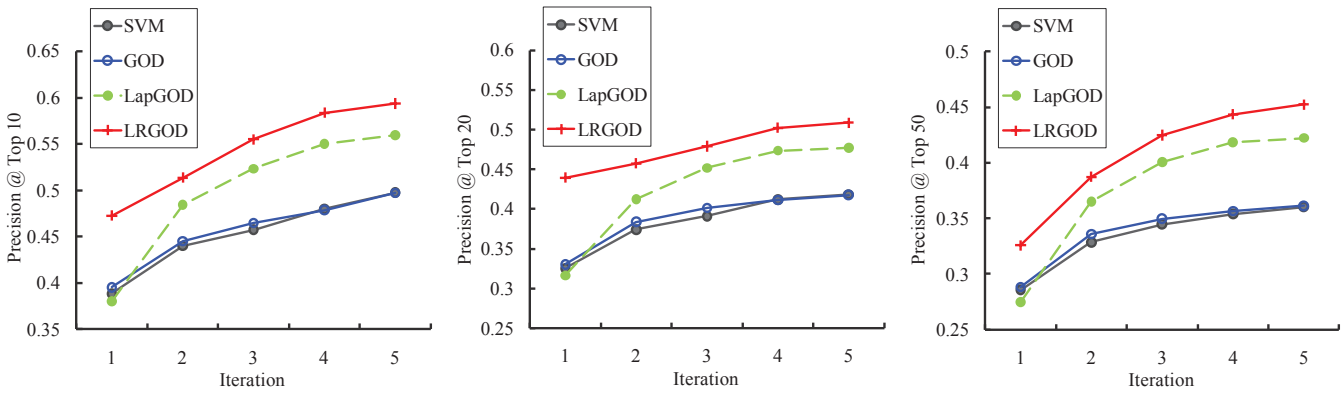

Figure 3: The precision rates of the four active learning approaches at Top 10, 20, 50 results on Corel dataset.

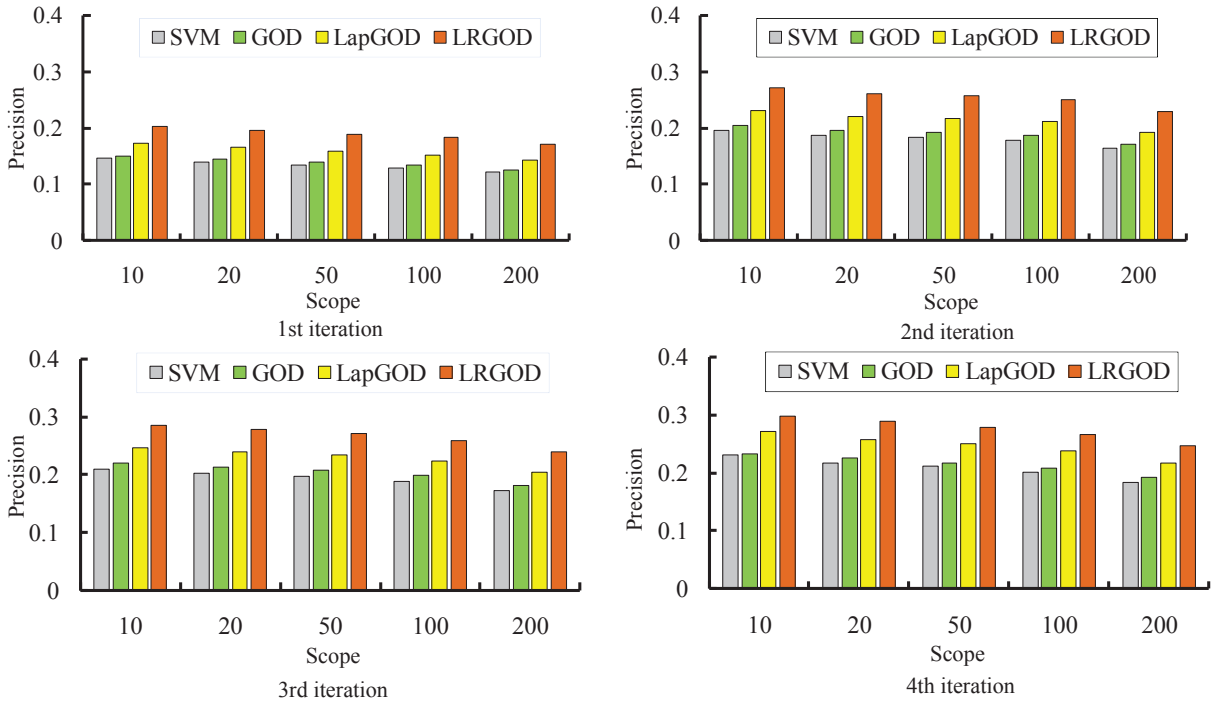

Figure 4: The performance of the four active learning approaches on NUS-WIDE-OBJECT dataset measured by Precision-Scope.

5.3 Experimental Results on NUS-WIDE-OBJECT Dataset

Figure 4 illustrates the precision-scope of the four active learning approaches in the first four iterations over the NUS-WIDE-OBJECT dataset. Figure 5 shows the precision rates at top 10, top 20, and 50 results in the five feedback iterations. From the experimental results, we can see that our LRGOD performs the best on the entire scope and in every iterations. Specifically, in the first iteration, it outperforms SVM*active*, GOD, and LapGOD by 38.6%, 34.7%, and 17.5%, in terms of precision at top 10 results, respectively. In the last iteration, it achieves 33.2%,27.9%,and 13.1% improvements at top 10 results, respectively. Consider top 20 results, LRGOD outperforms SVM*active*, GOD, and LapGOD by 39.7%, 34.8%, and 18.4% in the first iteration, respectively. In the last iteration, the improvements are 34.4%, 28%, and 13.9%, respectively.

6. CONCLUSIONS

In this paper, we have proposed a new active learning algorithm, called Locally Regressive G-Optimal Design (LRGOD) for relevance feedback image retrieval. Our algorithm is based on the assumption that for each image, its label values can be well estimated based on its neighbors via a locally regressive function. LRGOD is developed based on a locally regressive Least Squares model which makes use of both labeled and unlabeled images and simultaneously exploits the local structure of each image. It selects the most informative images to minimize the maximum predictive variance. Compared to existing active learning algorithms, such as the conventional Optimum Experimental Design and Laplacian Regularized Optimum Experimental Design, the proposed LRGOD approach makes use of data distribution by exploiting the local structure of each image. Therefore, the images selected by LRGOD can improve the retrieval performance the most, if they are used as training samples. We have evaluated the performance of LRGOD on two real-world image datasets: Corel and NUS-WIDE-OBJECT datasets. The experimental results validate that LRGOD achieves the best performance as compared to three representative active learning algorithms, and is robust to difference datasets.

The future works could focus on the following directions. First, in this paper, we use the G-Optimal Design criterion. The other classic optimal design criteria, such as D-, A-, E-, and I-Optimal Designs, can also be reformulated under this framework to exploit the local structure of each sample. Second, in some image retrieval scenarios, the relevant images are usually rare and hard to be found. Thus, it is useful to extend the proposed approach to tackle the insufficient relevant sample problem. Third, it is interesting to investigate the effectiveness of the proposed approach in other applications.

Acknowledgments

This research was supported by NRF (National Research Foundation of Singapore) Research Grant 252-300-001-490 under the NExT Search Center.

7. REFERENCES

[1] A. Atkinson, A. Donev, and R. Tobias. *Optimum experiment designs, with SAS*. Oxford University Press, 2007.

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

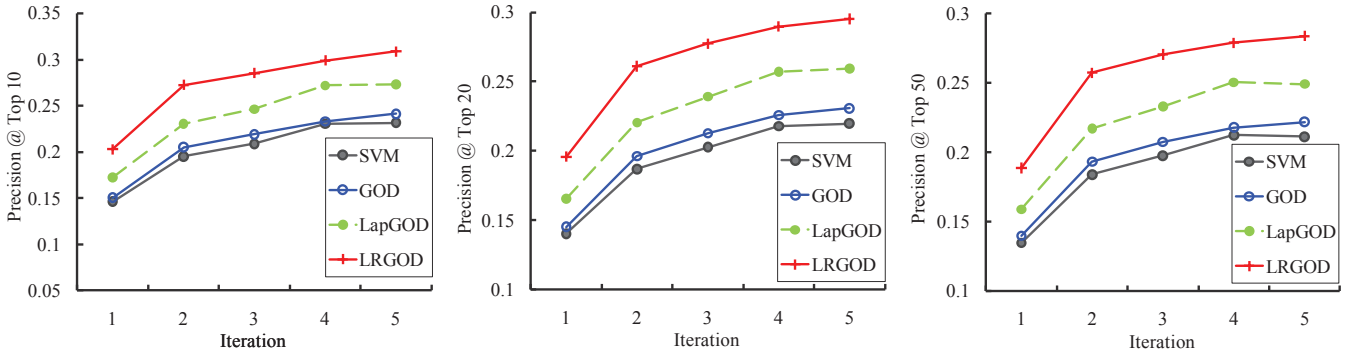

Figure 5: The precision rates of the four active learning approaches at Top 10, 20, 50 results on NUS-WIDE-OBJECT dataset.

[3] Chapelle, O., B. Schölkopf, and A. Zien. *Semi-supervised learning*. MIT Press, 2010.

[4] C. Chen, Z. Chen, J. Bu, C. Wang, L. Zhang, and C. Zhang. G-optimal design with laplacian regularization. In *Procedings of the AAAI Conference on Artificial Intelligence*, pages 254–263.

[5] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of ACM Conference on Image and Video Retrieval*, pages 1081–1088, Santorini, Greece, 2009.

[6] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: approaches and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[7] B. Geng, L. Yang, C. Xu, and X.-S. Hua. Content-aware ranking for visual search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[8] G. H. Golub and C. F. V. Loan. *Matrix Computations, 3rd Edition*. Johns Hopkins University Press, 1996.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2009.

[10] X. He. Laplacian regularized d-optimal design for active learning and its application to image retrieval. *IEEE Transactions on Image Processing*, 19(1):254–263, 2009.

[11] X. He, M. Ji, and H. Bao. A unified active and semi-supervised learning framework for image compression. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 65–72, Miami,USA, 2009.

[12] X. He, W. Min, D. Cai, and K. Zhou. Laplacian optimal design for image retrieval. In *Proceedings of the SIGIR conference on research and development in information retrieval*, pages 119–126, 2007.

[13] T. S. Huang, C. K. Dagli, S. Rajaram, E. Y. Chang, M. I. M. amd G. E. Polner, and D. P. W. Ellis. Active learning for interactive multimedia retrieval. *Proceeding of IEEE*, 96(4):648–667, 2008.

[14] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.

[15] T. Mei, Z.-J. Zha, and et al. Msra at trecvid 2008: High-level feature extraction and automatic search. *TRECVID Workshop*, 2008.

[16] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transaction on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

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

[18] M. Wang and X.-S. Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2(2), 2011.

[19] M. Wang, X.-S. Hua, J. Hong, R. Tang, G.-J. Qi, and Y. Song. Unified video annotation via multi-graph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(5), 2009.

[20] M. Wang, X.-S. Hua, J. Tang, and R. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Transactions on Multimedia*, 11(3), 2009.

[21] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang. Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia*, 12(8), 2010.

[22] M. Wu and B. Schölkopf. Transductive classification via local learning regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 628–635, Canada, 2007.

[23] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of the ACM International conference on Multimedia*, pages 175–184, New York, USA, 2009.

[24] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *Proceedings of the International Conference on Machine Learning*, pages 1081–1088, 2006.

[25] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label and multi-instance learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, USA, 2008.

[26] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, 20:97–103, 2009.

[27] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *Proceedings of the ACM international conference on Multimedia*, pages 15–24, Beijing,China, 2009.

[28] L. Zhang, C. Chen, W. Chen, J. Bu, D. Cai, and X. He. Convex experimental design using manifold structure for image retrieval. In *Proceedings of the ACM International conference on Multimedia*, pages 45–54, Beijing, China, 2009.

[29] Y.-T. Zheng, M. Zhao, S.-Y. Neo, and T.-S. Chua. Visual synset: Towards a higher-level visual representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, USA, 2008.

[30] X. Zhu. Semi-supervised learning literature survey. 2005.