

Convex Experimental Design Using Manifold Structure for Image Retrieval

Lijun Zhang¹
zljzju@zju.edu.cn

Chun Chen¹
chenc@zju.edu.cn

Wei Chen¹
chenw@zju.edu.cn

Jiajun Bu¹
bjj@zju.edu.cn

Deng Cai²
dengcai@cad.zju.edu.cn

Xiaofei He²
xiaofeihe@cad.zju.edu.cn

¹Zhejiang Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou, China

²State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou, China

ABSTRACT

Content Based Image Retrieval (CBIR) has become one of the most active research areas in computer science. Reverse feedback is often used in CBIR systems to bridge the semantic gap. Typically users are asked to make reverse judgments on some query results and the feedback information is then used to rerank the images in the database. An effective reverse feedback algorithm must provide the users with the most informative images with respect to the ranking function. In this paper we propose novel active learning algorithm based on Convex L₁ Regularized Optimal Design CLPRID for reverse feedback image retrieval. Our algorithm is based on regression model which initializes the least square error on the feedback images and simultaneously preserves the intrinsic geometric structure of the image space. It selects the most informative images which initialize the average predictive variance. The optimization problem of CLPRID can be cast as sequential programming SDP problem and solved via interior point methods. Experimental results on COREL dataset have demonstrated the effectiveness of the proposed algorithm for reverse feedback image retrieval.

Categories and Subject Descriptors

H Information storage and retrieval Information search and retrieval *Reverse feedback* G Mathematics of Computing Probability and Statistics *Experimental design*

General Terms

Algorithms Performance Theory

Keywords

Image retrieval active learning convex optimization reverse feedback sequential programming

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 fee.

MM '09, October 2009, Beijing, China
Copyright 2009 ACM 978-1-60558-848-8

1. INTRODUCTION

With the rapid increase in the volume of electronic archived image and video materials Content Based Image Retrieval (CBIR) has become one of the most active research areas for the last few decades. Query by example (QBE) is the traditional type of query in CBIR. In this environment users formulate query by means of giving negative feedback. CBIR systems use the observed visual features as a coarse representation and shape to represent the image's content and relevant images are retrieved based on the similarity of their visual features. Although CBIR has been extensively studied the semantic gap between observed features and high level semantic concepts is its performance gap.

To overcome the semantic gap reverse feedback is introduced into CBIR. Typically users are asked to make reverse judgments on the top images returned by the system and their preference is used to train classifier to separate images that match the query concept from those that do not. However in general the top returned images are not the most informative ones. In the worst case the top images are exactly the user's favorite and thus the standard classification techniques cannot be applied due to the lack of negative examples. The key problem then becomes how to select the most informative samples from the database. In machine learning this problem is called active learning which studies the phenomenon of learner selecting queries or asking queries that influence the training set.

Active learning algorithm is highly correlated with the underlying ranking mechanism. The most popular active learning techniques include Support Vector Machine active learning $S_{M_{active}}$ and regression based active learning R_{active} . $S_{M_{active}}$ asks the user to label those images which are closest to the $S_{M_{active}}$ boundary. The rationale is that the closer to the $S_{M_{active}}$ boundary an image is the less reliable its classification is. One of the major disadvantages of $S_{M_{active}}$ is that the estimated boundary may not occur enough especially when the number of feedback images is small. Moreover $S_{M_{active}}$ cannot be applied to the first round of the retrieval when there is no feedback images.

In statistics the problem of selecting samples to label is usually referred to sequential design. The study of Optimal Experimental Design (OED) is concerned with the design of experiments that are expected to minimize variance of parameterized model. There are two types of selection criteria of OED. One type is to choose points to minimize the confidence region for the estimated model parameters. High results in D- and E-optimal Design. The other is to minimize the variance of the pre-

dictionary \mathcal{V} which results in ℓ_1 and G optimization. Design Recent y
Yu et al. propose non-convex learning for ℓ_1 and G optimization. Conducted Transduc
tive Experiment Design TED 2 TED selects d that points to
minimize the average predictive variance of the learned function
on some pre-given dataset. It is so far used into convex optimization

TED is then formulated as the following optimization problem

$$\begin{aligned} \min \quad & \text{Tr}(X^T(ZZ^T + \gamma I)^{-1}X) \\ \text{s.t.} \quad & \{\mathbf{z}_1, \dots, \mathbf{z}_k\} \subseteq \mathcal{X} \end{aligned}$$

with variable $Z = [\mathbf{z}_1, \dots, \mathbf{z}_k]$. After solving the optimization problem, the above procedure can be reformulated as an equivalent optimization problem as follows

$$\begin{aligned} \min \quad & \sum_{i=1}^m \|\mathbf{x}_i - Z\boldsymbol{\alpha}_i\|^2 + \gamma \|\boldsymbol{\alpha}_i\|^2 \\ \text{s.t.} \quad & \{\mathbf{z}_1, \dots, \mathbf{z}_k\} \subseteq \mathcal{X} \end{aligned}$$

where the variables are $Z = [\mathbf{z}_1, \dots, \mathbf{z}_k]$ and $\boldsymbol{\alpha}_i \in \mathbb{R}^k, i = 1, \dots, m$

The above problem is NP-hard. Yu et al. have proposed sequential greedy algorithm and convex relaxation⁴² to solve it. The convex relaxation CovTED is shown as follows

$$\begin{aligned} \min \quad & \sum_{i=1}^m \left(\|\mathbf{x}_i - X\boldsymbol{\alpha}_i\|^2 + \sum_{j=1}^m \frac{\alpha_{i,j}^2}{\beta_j} \right) + \gamma \|\boldsymbol{\beta}\|_1 \\ \text{s.t.} \quad & \beta_j \geq 0, j = 1, \dots, m \end{aligned}$$

where the variables are $\boldsymbol{\beta} \in \mathbb{R}^m$ and $\boldsymbol{\alpha}_i \in \mathbb{R}^m, i = 1, \dots, m$. Here $\|\boldsymbol{\beta}\|_1$ is the ℓ_1 norm of $\boldsymbol{\beta}$ which is used to enforce some elements of $\boldsymbol{\beta}$ to be zero. An iterative algorithm is proposed to solve it⁴².

3. CONVEX LAPLACIAN REGULARIZED I-OPTIMAL DESIGN

Traditional active learning algorithms such as $S_{M_{active}}$ and OED are based on supervised learning algorithms S_{M} or linear regression. These approaches only consider the embedded points while neglecting the relationship of undetermined points which are very essential in selecting informative samples. We introduce in this section novel active learning algorithm which is based on one-step supervised learning algorithm. We first introduce the linear algorithm and then generalize it to the nonlinear case by applying kernel tricks. Our algorithm is fundamentally based on Laplacian Regularized Least Squares LapRLS and motivated by recent progress in perceptual design^{24, 242}.

3.1 Laplacian Regularized Least Squares

Laplacian Regularized Least Squares LapRLS makes use of other embedded undetermined points to discover the intrinsic geometric structure in the dataset. It assumes that if two points \mathbf{x}_i and \mathbf{x}_j are close then their responses $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ are close as well. Specifically LapRLS does not only preserve regularizer into the loss function of ridge regression Eq⁴. Let W be similarity matrix, the new loss function is defined as follows

$$J_L(\mathbf{w}) = \sum_i^k (f(\mathbf{z}_i) - y_i)^2 + \frac{\alpha}{2} \sum_{i,j=1}^m (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 W_{ij} + \beta \|\mathbf{w}\|^2 \quad 2$$

where $\alpha \geq 0$ and $\beta \geq 0$ are the regularization parameters. The second term of the right-hand side in the cost function is the locality preserving regularizer which incurs heavy penalty if neighboring points \mathbf{x}_i and \mathbf{x}_j are far apart.

There are many choices of similarity matrix W . A simple definition is as follows

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

Let D be diagonal matrix with $D_{ii} = \sum_{j=1}^m W_{ij}$ and $L = D - W$. The matrix L is called Graph Laplacian in spectral graph theory⁸. The solution to minimize equation 2 is given as follows

$$\hat{\mathbf{w}}_L = (ZZ^T + \alpha XLX^T + \beta I)^{-1} Z\mathbf{y} \quad 4$$

Let $H = ZZ^T + \alpha XLX^T + \beta I$ the covariance matrix of $\hat{\mathbf{w}}_L$ is

$$\begin{aligned} \text{Cov}(\hat{\mathbf{w}}_L) &= H^{-1} Z \text{Cov}(\mathbf{y}) Z^T H^{-1} \\ &= \sigma^2 H^{-1} Z Z^T H^{-1} \\ &= \sigma^2 H^{-1} (H - \alpha XLX^T + \beta I) H^{-1} \\ &= \sigma^2 H^{-1} - \sigma^2 H^{-1} (\alpha XLX^T + \beta I) H^{-1} \end{aligned}$$

Since the regularization parameters α and β are usually set to be very small, we have

$$\text{Cov}(\hat{\mathbf{w}}_L) \approx \sigma^2 H^{-1} = \sigma^2 (ZZ^T + \alpha XLX^T + \beta I)^{-1} \quad 5$$

3.2 Convex Laplacian Regularized I-optimal Design

Through making use of other embedded undetermined LapRLS estimates in the regularizing function $f(\mathbf{x}) = \hat{\mathbf{w}}_L^T \mathbf{x}$ that respects the intrinsic geometric structure in the dataset. An ideal design should choose subset $\mathcal{Z} \subseteq \mathcal{X}$ which simultaneously initializes the confidence region for $\hat{\mathbf{w}}_L^T$ and the predictive variance of $f(\mathbf{x})$. However, usually choice has to be made between these desiderata. In image retrieval, e.g., iterative regression function which can distinguish the relevant images from irrelevant ones. It is natural to require that the predictions of the learned function on the input set are as stable as possible. Thus, we use the I-optimal design criterion to select those images which can minimize the average predictive variance of learned regression function.

Here we consider set $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_l\}$ of test dataset points besides candidate set $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$. In specific, \mathcal{V} and \mathcal{X} can be the same set. Given test point \mathbf{v} , its prediction value is $f(\mathbf{v}) = \hat{\mathbf{w}}_L^T \mathbf{v}$ with variance $\text{Var}(f(\mathbf{v})) = \mathbf{v}^T \text{Cov}(\hat{\mathbf{w}}_L) \mathbf{v}$. Let $V = [\mathbf{v}_1, \dots, \mathbf{v}_l]$ the average predictive variance on \mathcal{V} is

$$\begin{aligned} & \frac{1}{l} \sum_{i=1}^l \mathbf{v}_i^T \text{Cov}(\hat{\mathbf{w}}_L) \mathbf{v}_i \\ & \approx \frac{\sigma^2}{l} \sum_{i=1}^l \mathbf{v}_i^T (ZZ^T + \alpha XLX^T + \beta I)^{-1} \mathbf{v}_i \quad 6 \\ & = \frac{\sigma^2}{l} \text{Tr}(V^T (ZZ^T + \alpha XLX^T + \beta I)^{-1} V) \end{aligned}$$

Then our problem is to find subset $\mathcal{Z} \subseteq \mathcal{X}$ to minimize equation 6. A simple sequential greedy approach is suggested to select \mathbf{z}_i 's one after another in⁴.

By introducing m indicator variables $\{\lambda_i\}_{i=1}^m \in \{0, 1\}$ here λ_i indicates whether or not point \mathbf{x}_i is chosen. Finding subset \mathcal{Z} to minimize equation 6 is equivalent to the following optimization problem

$$\begin{aligned} \min \quad & \text{Tr}(V^T (\sum_{i=1}^m \lambda_i \mathbf{x}_i \mathbf{x}_i^T + \alpha XLX^T + \beta I)^{-1} V) \\ \text{s.t.} \quad & \{\lambda_i\}_{i=1}^m \in \{0, 1\}, \sum_{i=1}^m \lambda_i = k \end{aligned} \quad 8$$

where the variables are $\{\lambda_i\}_{i=1}^m$ and k is the number of dataset points to be chosen. To simplify our presentation, we use vector $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_m]$ to denote the m variables. The variable vector $\boldsymbol{\lambda}$ is sparse and has only k non-zero entries.

nonnegative values. Then the value of λ_i indicates how significant \mathbf{x}_i contributes to the minimization in problem 8. The sparseness of λ can be controlled through minimizing the ℓ_1 norm of λ which is a very popular technique in regression.⁴

Following the convention in the field of optimization we use $\lambda \succeq 0$ to denote that the elements in λ should be nonnegative. And we use the elements of λ be nonnegative $\|\lambda\|_1$ is equal to $\mathbf{1}^T \lambda$ here $\mathbf{1}$ is column vector containing ones. Finally our optimization problem becomes

Definition Convex Lp- α Regularized I-optimal Design CL-pRID

$$\text{in } \text{Tr}(V^T (\sum_{i=1}^m \lambda_i \mathbf{x}_i \mathbf{x}_i^T + \alpha X L X^T + \beta I)^{-1} V) + \gamma \mathbf{1}^T \lambda$$

st $\lambda \succeq 0$

here the variable is $\lambda \in \mathbb{R}^m$ and $\gamma \geq 0$ is the trade-off parameter for sparsity

THEOREM *Problem 9 is convex optimization problem with variable $\lambda \in \mathbb{R}^m$*

PROOF Let $g(X) = \text{Tr}(V^T X^{-1} V) = \sum_{j=1}^l \mathbf{v}_j^T X^{-1} \mathbf{v}_j$ and $h(\lambda) = \sum_{i=1}^m \lambda_i \mathbf{x}_i \mathbf{x}_i^T + \alpha X L X^T + \beta I$ we know that the trace function $f_1(X) = \mathbf{v}^T X^{-1} \mathbf{v}$ is convex function of X .⁴ Since nonnegative weighted sum preserves convexity $g(X)$ is so convex function of X we define

$$g \circ h(\lambda) = \text{Tr}(V^T (\sum_{i=1}^m \lambda_i \mathbf{x}_i \mathbf{x}_i^T + \alpha X L X^T + \beta I)^{-1} V)$$

Because $h(\lambda)$ is affine function of λ and composition with affine function preserves convexity $g \circ h$ is convex function of λ

Since $f_2(\lambda) = \gamma \mathbf{1}^T \lambda$ is convex function of λ the objective function of problem $g \circ h(\lambda) + f_2(\lambda)$ is so convex

Because the objective function is convex the inequality constraint function $-\lambda$ is convex problem is convex optimization problem with variable $\lambda \in \mathbb{R}^m$.⁴ \square

3.3 Optimization Scheme

The success of Semidefinite programming (SDP) in various applications motivates us to formulate and solve CL-pRID as an SDP problem. Semidefinite programming has seen the most exciting theoretical development in the topic programming. It has applications in traditional convex constrained optimization systems in such diverse domains as control theory and combinatorial optimization.² Moreover the powerful interior point methods for linear programming have been extended to SDP.

By introducing the variable $P \in \mathbb{R}^{l \times l}$ optimization problem can be equivalently reformulated as

$$\text{in } \text{Tr}(P) + \gamma \mathbf{1}^T \lambda$$

st $P \succeq_{S_{n+l}^+} V^T (\sum_{i=1}^m \lambda_i \mathbf{x}_i \mathbf{x}_i^T + \alpha X L X^T + \beta I)^{-1} V$ 2

$\lambda \succeq 0$

with variables $P \in \mathbb{R}^{l \times l}$ and $\lambda \in \mathbb{R}^m$. Here S_{n+l}^+ denotes the set of symmetric positive semidefinite $(n+l) \times (n+l)$ matrices which is called positive semidefinite cone in the field of optimization. The associated generalized inequality $\succeq_{S_{n+l}^+}$ is the usual matrix inequality $A \succeq_{S_{n+l}^+} B$ means $A - B$ is positive semidefinite $(n+l) \times (n+l)$ matrix.⁴

THEOREM *Problem 9 is equivalent to problem 2*

PROOF Let λ_a^* be the optimal solution of problem 9 and P^*, λ_b^* be the optimal solutions of problem 2. Then λ_a^*, λ_b^* is

sufficient condition for Theorem 2. Let $f(\lambda) = \text{Tr}(V^T (\sum_{i=1}^m \lambda_i \mathbf{x}_i \mathbf{x}_i^T + \alpha X L X^T + \beta I)^{-1} V)$

Assume $\lambda_a^* \neq \lambda_b^*$. Since λ_a^* minimizes problem 9 we must have $\text{Tr}(f(\lambda_a^*)) + \gamma \mathbf{1}^T \lambda_a^* < \text{Tr}(f(\lambda_b^*)) + \gamma \mathbf{1}^T \lambda_b^*$. Because P^*, λ_b^* satisfies the constraints in problem 2 we have

$$\begin{aligned} P^* \succeq_{S_l^+} f(\lambda_b^*) &\Leftrightarrow P^* - f(\lambda_b^*) \in S_l^+ \\ &\Rightarrow \text{Tr}(P^* - f(\lambda_b^*)) \geq 0 \\ &\Rightarrow \text{Tr}(P^*) \geq \text{Tr}(f(\lambda_b^*)) \\ &\Rightarrow \text{Tr}(P^*) + \gamma \mathbf{1}^T \lambda_b^* \geq \text{Tr}(f(\lambda_b^*)) + \gamma \mathbf{1}^T \lambda_b^* \\ &\Rightarrow \text{Tr}(P^*) + \gamma \mathbf{1}^T \lambda_b^* > \text{Tr}(f(\lambda_a^*)) + \gamma \mathbf{1}^T \lambda_a^* \end{aligned}$$

It is clear that $(f(\lambda_a^*), \lambda_a^*)$ satisfies the constraints in problem 2. Then for problem 2 $(f(\lambda_a^*), \lambda_a^*)$ is more optimal than P^*, λ_b^* which contradicts our assumptions. So we must have $\lambda_a^* = \lambda_b^*$. \square

Problem 2 can be cast as an SDP using the Schur complement theorem.⁴ Given symmetric tripartitioned

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

If A is invertible the matrix $S = C - B^T A^{-1} B$ is called the Schur complement of A in X . Schur complement theorem states that if A is positive definite then X is positive semidefinite if and only if S is positive semidefinite.

According to this theorem problem 2 is equivalent to the following semidefinite programming SDP

$$\text{in } \text{Tr}(P) + \gamma \mathbf{1}^T \lambda$$

st $\begin{bmatrix} \sum_{i=1}^m \lambda_i \mathbf{x}_i \mathbf{x}_i^T + \alpha X L X^T + \beta I & V \\ V^T & P \end{bmatrix} \succeq_{S_{n+l}^+} 0$ 2

$\lambda \succeq 0$

with variables $P \in \mathbb{R}^{l \times l}$ and $\lambda \in \mathbb{R}^m$. As explained previously $A \succeq_{S_{n+l}^+} 0$ means A is positive semidefinite $(n+l) \times (n+l)$ matrix. Hence so we can solve this problem exactly via interior point methods.⁴ After obtaining the optimal solution λ^* we select k points with the largest significant indicators λ_i^* for user to use.

4. CONVEX KERNEL LAPLACIAN REGULARIZED I-OPTIMAL DESIGN

In traditional point design on \mathcal{Y} considers linear functions when the data is highly nonlinear distributed the linear function might not be able to fit the data. In this section we extend CL-pRID to handle nonlinear cases by performing equivalent design in the Reproducing Kernel Hilbert Space (RHS). We begin with brief description of kernel Lp- α Regularized Least Squares

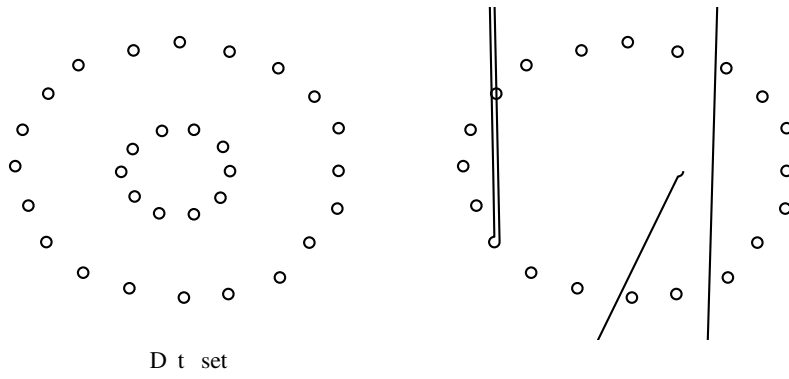
4.1 Kernel Laplacian Regularized Least Squares

Let K be positive definite Mercer kernel $K: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ and \mathcal{H}_K be the corresponding Reproducing Kernel Hilbert Space (RHS). Consider the optimization problem 2 in RHS. Then we seek function $f \in \mathcal{H}_K$ such that the following objective function is minimized

$$J_L(f) = \sum_{i=1}^k (y_i - f(\mathbf{z}_i))^2 + \frac{\alpha}{2} \sum_{i,j=1}^m (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 W_{ij} + \beta \|f\|_{\mathcal{H}_K}^2$$

The Representer Theorem can be used to show that the solution is an expansion of kernel functions over both the training and the unlabeled data

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^m \hat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i)$$

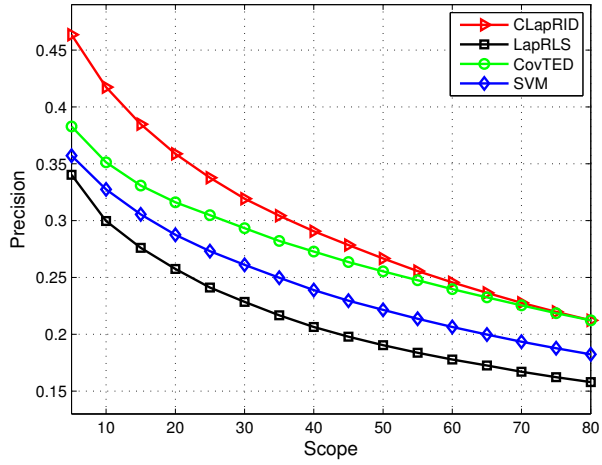


precision-scope cur e

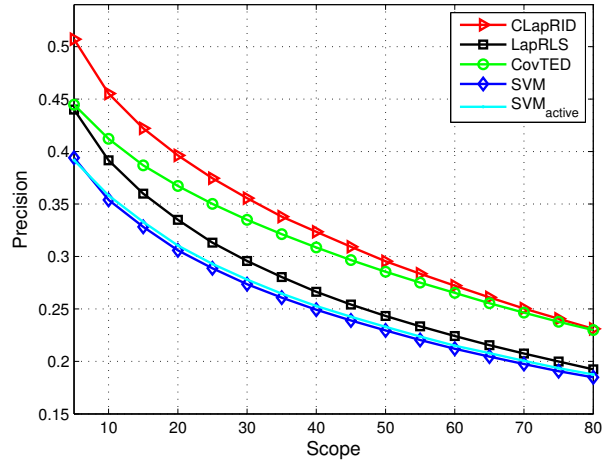
eeperi ent design eo

ithms. The numbers beside the selected points denote their o

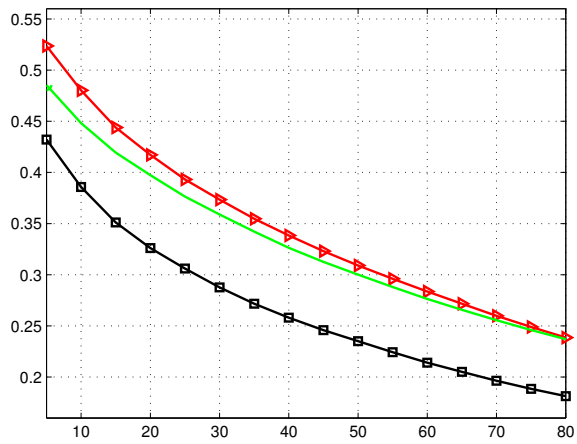
o p r isons ith other gorith s e



First Iteration



Second Iteration



rather even decreases after the second feedback iteration. This phenomenon vividly shows that the top images are not the most informative ones.

6. CONCLUSIONS

In this paper we propose a novel active learning algorithm called Convex L_p constrained Regularized Iterative Design CL_pIRID for relevance feedback image retrieval. Our algorithm is fundamentally based on L_p constrained Regularized Least Squares L_pRLS and motivated by many recent advances in efficient design [24].

CL_pIRID makes use of other established and unexplored points to discover the intrinsic geometric structure in the data. It selects images to minimize the variance of prediction value and can also avoid side effects by programming Efficient Results on COREL dataset. It shows that the proposed approach outperforms Support Vector Machines L_p constrained Regularized Least Squares Support Vector Machine Active Learning [2]. Convex Truncative Efficient Design [4].

In this paper we use Iterative design criterion. However other classic optimization criteria such as D₁, A₁, E₁ and G₁ optimization designs can also be reformulated under this framework to reflect the underlying geometric structure.

7. ACKNOWLEDGMENTS

This work is supported by China National Key Technology R & D Program [2008BAH1B01201, 2008BAH1B01202].

8. REFERENCES

- [1] S. P. Asprey and S. M. Chietto, "Designing robust optimization criteria for dynamic systems," *Journal of Process Control*, vol. 14, no. 2, pp. 101-110, 2004.
- [2] A. Atkinson, A. Donev, and R. Tobias, *Optimal Experimental Designs with SAS*. Oxford University Press, USA, 2009.
- [3] M. Bekin, P. Niyogi, and S. Sridharan, "Minimizing the variance of prediction value and avoiding side effects in active learning," *The Journal of Machine Learning Research*, vol. 10, pp. 101-110, 2009.
- [4] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [5] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.
- [6] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *Software: Practice and Experience*, vol. 32, no. 4, pp. 1365-1373, 2002.
- [7] O. Chapelle, B. Schölkopf, and A. Zien, editors, *Support Vector Machines*. MIT Press, Cambridge, MA, 2000.
- [8] F. R. Chung, *Spectral Graph Theory*. volume 9 of *Regional Conference Series in Mathematics*. American Mathematical Society, 1997.
- [9] D. A. Cohen, Z. Ghahramani, and M. I. Jordan, "Active and passive learning with stochastic models," *Journal of Artificial Intelligence Research*, vol. 10, pp. 1-35, 2000.
- [10] R. D. D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences and trends of the near future," *ACM Computing Surveys*, vol. 42, no. 2, pp. 1-51, 2009.
- [11] E. Dekkers, *Aspects of Sequential Programming*. 2002. -2- - Td R Pcngrthth onite ertsd Appoig rinnf