Manifold Learning using Robust Graph Laplacian for Interactive Image Search

Hichem Sahbi
UMR 5141, CNRS
Telecom ParisTech, France
sahbi@telecom-paristech.fr

Patrick Etyngier, Jean-Yves Audibert and Renaud Keriven
Certis Lab
ENPC ParisTech, France
{etyngier,audibert,keriven}@certis.enpc.fr

Abstract

Interactive image search or relevance feedback is the process which helps a user refining his query and finding difficult target categories. This consists in partially labeling a very small fraction of an image database and iteratively refining a decision rule using both the labeled and unlabeled data. Training of this decision rule is referred to as transductive learning.

Our work is an original approach for relevance feedback based on Graph Laplacian. We introduce a new Graph Laplacian which makes it possible to robustly learn the embedding, of the manifold enclosing the dataset, via a diffusion map. Our approach is three-folds: it allows us (i) to integrate all the unlabeled images in the decision process (ii) to robustly capture the topology of the image set and (iii) to perform the search process inside the manifold. Relevance feedback experiments were conducted on simple databases including Olivetti and Swedish as well as challenging and large scale databases including Corel. Comparisons show clear and consistent gain, of our graph Laplacian method, with respect to state-of-the art relevance feedback approaches.

1. Introduction

At least, two interrogation modes are known in content based image retrieval (CBIR); the query by example and relevance feedback (RF). In the first mode the user submits a query image as an example of his "class of interest" and the system displays the closest image(s) using a feature space and a suitable metric [5, 25]. A slight variant is category retrieval which consists in displaying images belonging to the "class of the query". In the second category (see the pioneering works [14, 19]) the user labels a subset of images as positive and/or negative according to an unknown metric defined in "his mind" and the CBIR system refines a metric and/or a decision rule and displays another set of images hopefully closing the gap between the user's intention and the response(s) of the the CBIR system [31, 8, 20, 32]. This process is repeated until the system converges to the user's class of interest. The performance of an RF system is usually measured as the expectation of the number of user(s)' responses (or iterations) necessary to focus on the targeted class. This performance depends on the capacity of an RF system (i) to generalize well on the set of unlabeled images using the labeled ones, (ii) to ask the most informative questions to the user (see for instance [28]) and (iii) the consistency and self-consistency of the user(s)' responses. Points (i)–(ii) are respectively referred to as *the transduction* and the *display models*. Point (iii) assumes that different users have statistically the same answers according to an existing but unknown model referred to as the *user model*.

1.1. Related Work

Different schemes exist in the literature for the purpose of RF [20, 32] which are either based on density estimation [17, 13] or discriminative training [28], depending respectively on the fact that they model the distribution and *the topology* of the positive and (possibly) the negative labeled images or they build a decision function which classifies the unlabeled data. In the first category, different density estimation methods are used in RF including non parametric Parzen windows [17], Gaussian mixture models [8], logistic regression [6] and novelty detectors [7, 23]. In [8, 9], the authors introduced a notion of relative judgment of the user, i.e., the response is not binary but a relative number measuring the relevance of a displayed set of images. The user's response is assumed as a sigmoid function of the distance, so images close to the highly numbered set are more likely to be the target than the others. The authors in [8] used Gaussian mixture models and a Bayesian framework in order to estimate (and update) a distribution through all images and display those which the highest probability. The proposed approach in [9] defines a criteria based on the mutual information between the user's response and all the possible target images in the database and display those which maximize this criteria.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

In the second family, discriminative methods learn from the aggregated set of positive and negative labeled images how to classify the unlabeled ones. Existing RF methods use support vector machines [28, 28, 10], decision trees [16], boosting [27] and Bayesian classifiers [10, 29, 8]. The RF method in [28] shows a particular interest by its important gain in the convergence speed when using active learning [22, 2].

1.2. Motivation and Contribution

The success of relevance feedback is largely dependent on how much (1) the image description (feature+similarity) fits (2) the semantic wanted by the user. The gap between (1) and (2) is referred to as *the semantic gap*. The reduction of this gap basically requires adapting the decision rule (as discussed earlier) and the features to the user's feedback. Many works (see, for instance [20]) consider features as a weighted combination of simple sub-features each one captures a particular characteristic. The weight of each sub-feature and hence the topology of the manifold enclosing the data is adapted by taking into account the variance of the labeled set, so relevance feedback will pay more attention to the sub-features with high variances. Put differently, adapting features might be explicitly achieved as in [20] or implicitly as a part of the decision rule training (as discussed in Section 1.1).

When the original sub-features are highly correlated, it is difficult to find dimensions, in the original feature space, which are clearly discriminant according to the user's feedback. This follows when the Gaussian assumption (about the distribution of the data) does not hold or when the classes are highly not separable, i.e., the data in original feature space form a non-linear manifold (see Figure 1, left). Therefore, further-processing is required in order to extract dimensions with high intrinsic variances. A didactic example, shown in Figure (1), (the application is searching faces by identity), follows the statement in [1]: *the variance due to the intra-class variability (pose, illumination, etc.) is larger than the inter-class variability (identity)*. Figure (1) illustrates this principle where clearly the intra-class variance estimated through the original feature space (resp. the intrinsic dimensions of the manifold enclosing the data) is larger (resp. smaller) than the inter-class variance. *Clearly, searching those faces through the intrinsic dimensions of the manifold is easier than in the original space.* Hence, learning the manifold enclosing the data is crucial in order to capture the *actual* topology of the data.

In this paper, we introduce a new relevance feedback scheme based on graph Laplacian[4]. We first model the topology of the image database, including the unlabeled images, using an eigen approximation of the graph Laplacian,

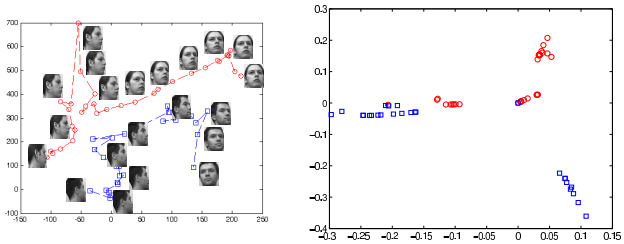

Figure 1. (Left) This figure shows the distribution of two classes corresponding to two individual. It is clear that the intra class variance is larger than the inter class one. (Right) This is the distribution of the same classes inside the manifold trained using graph Laplacian. It is clear that the converse is now true and the classification task is easier in the embedding space.

then we propagate the labels by projecting the whole dataset using a linear operator learned on both the labeled and the unlabeled sets. The main contributions of this work are:

- In contrast to existing relevance feedback methods which only rely on the labeled set of images, our approach integrates the unlabeled data in the training process through the cluster assumption [24] (As discussed in Section 3.1). These unlabeled data turn out to be very useful when only few labeled images are available since it allows us to favor decision boundaries located in low density regions of the image database, which are very often encountered in practice. The approach even though proved to work in the particular task of relevance feedback, it can be easily extended to other transductive learning tasks such as database categorization.

- In the second main contribution of this work, we derive a new from of the graph Laplacian which makes it possible to embed the dataset in *a robust way*. This graph Laplacian, based on diffusion map, captures the conditional probabilities of transition from any sample to another with a path of a given length. Its particularity is to only consider the intermediate paths with high transition likelihoods (see Section 3.2).

- For numerical and practical matters, we show in Section (4) the extension of the method in order to handle large scale databases using Nyström interpolation.

In the remainder of this paper, we consider the following notation. X is a random variable standing for a training sample taken from $\mathcal{X}$ and Y its class label in $\{+1, -1\}$ ($Y = 1$ if the sample X belongs to the targeted class and -1 otherwise). $G = \langle V, E \rangle$ denotes a graph where V is a set of vertices and E are weighted edges. We use also l, t as indices for iterations. Among terminologies a *display* is a set of images taken from the database which are shown to the user at iteration t. The paper is organized as

follows: Section 2 introduces the overall architecture of the RF process. Section 3 describes our RF model based on the s-weighted robust graph Laplacian and the display model. Section 4 provides an extension of the embedding method in order to handle large scale databases which are very often encountered in practice, using the Nyström operator. Section 5 provides an extensive experimental study using different databases including specific ones; face databases and also generic databases. We discuss the method and we conclude in Section 6.

2. Overview of the Search Process

Let $\mathcal{S} = \{X_1, ..., X_n\}$, $\mathcal{Y} = \{Y_1, ..., Y_n\}$ denote respectively a training set of images and the underlying unknown ground truth. Here Y_i is equal $+1$ if the image X_i belongs to the user's "class of interest" and $Y_i = -1$ otherwise. Let us consider $\mathcal{D}_t \subset \mathcal{S}$ as a display shown at iteration t and $\mathcal{Y}_t$ the labels of $\mathcal{D}_t$. Our interaction consists in asking the user questions such that his/her responses make it possible to reduce the *semantic gap* according to the following steps:

- "Page Zero": Select a display $\mathcal{D}_1$ which might be a random set of images or the prototypes found after applying clustering or Voronoi subdivision.
- Reduce the "semantic gap" iteratively ($t = 1$,..., T):

  (1) Label the set $\mathcal{D}_t$ using a (possibly stochastic) *known-only-by-the-user* function $\mathcal{Y}_t \leftarrow \mathcal{L}(\mathcal{D}_t)$. Here $\mathcal{L}$ is referred to as the user model which, given a display $\mathcal{D}_t$, provides the labels $\mathcal{Y}_t$. When the ground truth is unique, this function is consistent (through different users) and self-consistent (with respect to the same user) so the user's answer is coherent and objective, otherwise the labeling function becomes stochastic. The coherence issue is not in the scope of this paper (see [9] for a comprehensive study), so we only consider consistent and self-consistent users.

  (2) Train a decision function $f_t : \mathcal{X} \rightarrow \{-1, +1\}$ on the (so far) labeled training set $\mathcal{T}_t = \bigcup_{l=1}^{t}(\mathcal{D}_l, \mathcal{Y}_l)$ and the unlabeled set of images $\mathcal{S} - \cup_{l=1}^{t}\mathcal{D}_l$. The transduction model discussed in (3.1) is the one used for this training. At iteration t, the target is to efficiently use both labeled and unlabeled data in order to estimate the actual decision function,

$$\underset{f:\mathcal{X}\rightarrow\{+1;-1\}}{\operatorname{argmin}} \; P\left[f\left(X\right) \neq Y\right]. \tag{1}$$

  In our setting, it is important to generalize well even when the size of the labeled training set is small. This is why this step should use transductive methods which implicitly assume that the topology of the decision boundary depends on the unlabeled set $\mathcal{S} - \cup_{k=1}^{t}\mathcal{D}_k$ as shown in (3). More precisely, the clustering assumption implicitly made is: the decision boundary is likely to be in low density regions of the input space $\mathcal{X}$ [18]. (3) Select the next display $\mathcal{D}_{t+1} \subset \mathcal{S} - \bigcup_{k=1}^{t}\mathcal{D}_k$. The convergence of the RF model to the actual decision boundary is very dependent on the amount of information provided by the user. As $P(.)$ is unknown and the the whole process is computationally expensive, the display model considers a sampling strategy which selects a collection of images that improves our current estimate of the "class of interest" (see Section 3.3). This can be achieved by showing samples of difficult-to-classify images such as those close to the decision boundary. Given the labeled set $\mathcal{T}_t$, and let $f_{\mathcal{D}}$ be a classifier trained on $\mathcal{T}_t$ and a display $\mathcal{D}$. The issue of selecting $\mathcal{D}_{t+1}$ can be formulated at iteration $t+1$ as:

$$\begin{array}{l} \mathcal{D}_{t+1} \leftarrow \underset{\mathcal{D}}{\operatorname{argmin}} \;\; P\left[f_{\mathcal{D}}(X) \neq Y\right] \\ s.t. \;\; \mathcal{D}_{t+1} \bigcap \left(\bigcup_{l=1}^{t}\mathcal{D}_l\right) = \emptyset \end{array} \tag{2}$$

3. Graph Laplacian and Relevance Feedback

Graph Laplacian methods emerged recently as one of the most successful in transductive inference [4], (spectral) clustering [26] and dimensionality reduction [3]. The underlying assumption is: *the probability distribution generating the (input) data admits a density with respect to the canonical measure on a sub-manifold of the Euclidean input space*. Let $\mathcal{M}$ denotes this sub-manifold and p the probability distribution of the input space with respect to the canonical measure on $\mathcal{M}$ (i.e. the one associated with the natural volume element dV). Note that $\mathcal{M}$ can be all the Euclidean space (or a subset of it of the same dimension) so that p can simply be viewed as a density with respect to the Lebesgue measure on the Euclidean space.

3.1. s-Weighted Transductive Learner

In transductive inference, one searches for a smooth function $f : \mathcal{X} \rightarrow \mathcal{Y}$ from the input feature space into the output space such that $f(X_i)$ is close to the associated output Y_i on the training set and such that the function is allowed to vary only on low density regions of the input space. Let $s \geq 0$ be a parameter characterizing how low the density should be to allow large variations of f (see (3)). Depending on the confidence we assign to the training outputs, we obtain the following optimization problem:

$$\min_{f} \; \sum_{i=1}^{n} c_i \left[Y_i - f(X_i)\right]^2 + \int_{\mathcal{M}} \|\nabla f\|^2 \, p^s \, dV, \tag{3}$$

where the c_i's are positive coefficients measuring how much we want to fit the training point (X_i, Y_i). Typically, $c_i = +\infty$ imposes a hard constraint on the function f so that $f(X_i) = Y_i$. The s-th weighted Laplacian operator is characterized by:

$$\int_{\mathcal{M}} f \times (\Delta_s g)\, p^s dV = \int_{\mathcal{M}} \langle \nabla f, \nabla g \rangle\, p^s dV,$$

where f, g are infinitely smooth real-valued functions defined on $\mathcal{M}$ with compact support. By the law of large numbers, the integral in (3) can then be approximated by

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) \Delta_s f(X_i) p^{s-1}(X_i). \tag{4}$$

Unfortunately, the direct computation of $\Delta_s f(X_i)$ for every possible function f is not possible and solving (3) is intractable. *A discrete approximation of the s-th weighted Laplacian operator, proposes an alternative to this problem.* The method is based on a neighborhood graph in which the nodes are the input data from both the labeled and unlabeled sets. Again, let $X_1, \ldots, X_n$ denote these data and let $\tilde{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetrical non-negative function giving the similarity between two input points. The typical kernel is the Gaussian $\tilde{K}(x', x'') = e^{-\frac{\|x'-x''\|^2}{2\sigma^2}}$ and its degree function is defined as $\tilde{d}(x) = \sum_{i=1}^{n} \tilde{K}(X_i, x)$. Let $\lambda \geq 0$. General graph Laplacian methods use the normalized kernel defined as

$$K(x', x'') = \frac{\tilde{K}(x', x'')}{[\tilde{d}(x')\tilde{d}(x'')]^{\lambda}}. \tag{5}$$

Similarly, the degree function associated with the kernel K is $d(x) = \sum_{i=1}^{n} K(X_i, x)$. The kernel K induces a weighted undirected graph G in which the nodes are $X_1, \ldots, X_n$ and in which any two nodes are linked with an edge of weight $K(X_i, X_j)$. The degree of a node is defined by the sum of the weights of the edges at the node, i.e. $d(X_j)$.

Let W be the $n \times n$ matrix in which the generic element is $K(X_i, X_j)$. Let D be the diagonal $n \times n$ matrix for which the i-th diagonal element is $d(X_i)$. The matrix $L_{\text{rw}} = D^{-1}W$ defines the random walk graph Laplacian where the entry at row i and column j characterizes the probability of a walk from the node X_i to X_j. For a given $f : \mathcal{X} \to \mathcal{Y}$, let F be the vector defined as $F_i = f(X_i)$. The main result of [12] is essentially $L_{\text{rw}} F)_i \rightsquigarrow \big(\Delta_{2(1-\lambda)} f\big)(X_i)$ where $\rightsquigarrow$ means convergence almost sure when the sample of size n goes to infinity and the kernel bandwidth h goes to zero not too rapidly (e.g. $h = (\log n)^{-1}$), up to normalization of the left-hand side by an appropriate function of n and h. Besides one can understand the role of the degree functions through the convergences: $\tilde{d}(x) \rightsquigarrow p(x)$, and $d(x) \rightsquigarrow [p(x)]^{1-2\lambda}$. The above analysis shows that instead of focusing on the intractable optimization (3), one should solve its discrete counterpart for $\lambda = 1 - s/2$ in (5):

$$\min_{F \in \mathbb{R}^n} \; (F - Y)^t C (F - Y) + F^t \tilde{L} F,$$

whose solutions are of the linear system $(\tilde{L} + C)F = CY$ where $\tilde{L} = D^{s-1} L_{\text{rw}}$ in view of (4). Here C is the diagonal $n \times n$ matrix for which the i-th diagonal element is c_i for a labeled point, and 0 for an unlabeled point, and similarly, Y is the n-dimensional vector for which the i-th element is Y_i for a labeled point, and 0 for an unlabeled point.

3.2. Our Robust k-step Graph Laplacian

Let us rewrite L_{rw} as L. When embedding a dataset using the one step random walk graph Laplacian L, the main drawback is its sensitivity to noise. This comes from short-cuts, when building the adjacency graph (or estimating the scale parameter of the Gaussian kernel). Therefore, the *actual* topology of the manifold $\mathcal{M}$ will be lost (see. Figure 2, left). In [15], the authors consider instead a diffusion map graph Laplacian $L^{(k)}$ (denoted also L_k), here $L_k = L_{k-1} \times L$. The latter models a Markovian process where the conditional k-step transition likelihood (between two data X_i and X_j) is the sum of the conditional likelihoods of all the possible (k-1)-steps linking X_i and X_j. This results into low transition probabilities in low density areas. Nevertheless, when those areas are noisy, the method fails in capturing the correct topology (cf. Figure 2, middle).

Our k-step graph-Laplacian: the above limitation motivates the introduction of a new (called robust) graph Laplacian[1], recursively defined as

$$L_k \;=\; [L_{k-1}^{\frac{1}{\alpha}} \times L^{\frac{1}{\alpha}}]^{\alpha}, \quad 1/\alpha \in [1, +\infty[\tag{6}$$

Let $L(i,j)^{\frac{1}{\alpha}}$ denotes the j^{th} column of the i^{th} row of $L^{\frac{1}{\alpha}}$. Again, L is the one step random walk graph Laplacian where each entry $L(i,j)$ corresponds to the probability of a walk from X_i to X_j in one step, also denoted $P_1(j|i)$. This quantity characterizes the first order neighborhood structure of the graph G. In the context of diffusion map[15], the idea is to represent higher order neighborhood by taking powers of the matrix L, so $L_k(i,j) = P_k(j|i)$ will be the probability of a walk from X_i to X_j in k steps. Here k acts as a scale factor and makes it possible to increase the local influence of each node in the graph G. The matrix L_k can be inferred from L_{k-1} and L by summing the conditional probabilities

[1] Without any confusion and in the remainder of this paper, we denote by L_k this new form of the graph Laplacian.

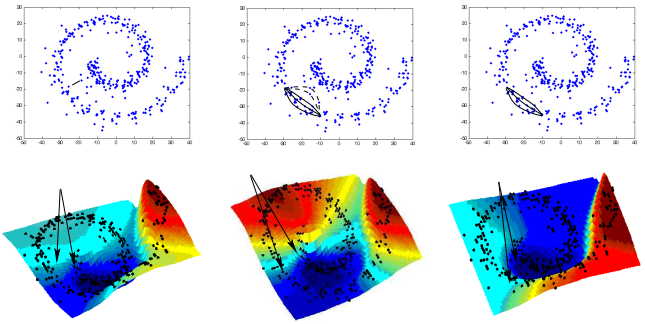

Figure 2. The top figures show samples taken from the Swiss roll. (left) A short cut makes the random walk Laplacian embedding very noise sensitive, clearly the variation of the color map does not follow the intrinsic dimension of the actual manifold. (middle) When using the diffusion map, noisy paths affect the estimation of the conditional probabilities. This issue is overcome in (right) when using the robust diffusion map, as now the color map varies following the intrinsic dimension.

over different paths , i.e.,

$$[P_k(j|i)]^{\frac{1}{\alpha}} = \sum_{l=1}^{n} [P_{k-1}(l|i)]^{\frac{1}{\alpha}} \, [P_1(j|l)]^{\frac{1}{\alpha}} \qquad (7)$$

We refer to a k-path as any path of k steps in the graph G. Depending on α the general form of the graph Laplacian L_k implements the following random walks:

- $\alpha \to 1$: $[P_k(j|i)]^1$ is *the average* transition probability of the k-paths linking X_i to X_j. So L_k implements exactly the one in [15].
- $\alpha \to 0$: it is easy to see that $[P_k(j|i)]^{\frac{1}{\alpha}}$ converges to $\max_l \{[P_{k-1}(l|i)]^{\frac{1}{\alpha}} \, [P(j|l)]^{\frac{1}{\alpha}}\}$, so $L_k(i,j)$ corresponds to *the most likely* transition probability of k-steps.
- $\alpha \in]0,1[$: $[P_k(j|i)]^{\frac{1}{\alpha}}$ is dominated by the largest terms in $\{[P_{k-1}(l|i)]^{\frac{1}{\alpha}} \, [P(j|l)]^{\frac{1}{\alpha}}\}$. The effect of noisy terms will then be reduced.

Figure (2, right) shows the application of (6) in the embedding of the Swiss roll data ($k = 10$ and $\alpha = 0.2$). Clearly, the topology of the data is now preserved. Figure (3) shows the robustness of the method to different amount of noise (again $k = 10$ and $\alpha = 0.2$).

3.3. Display Model

The data in $\mathcal{S}$ are mapped into a manifold $\mathcal{M}$ such that any two elements X_i and X_j in $\mathcal{S}$ with close conditional probabilities $\{P_k(i|.)\}$ and $\{P_k(j|.)\}$ will also be close in $\mathcal{M}$. Let Λ be the diagonal matrix of positive eigenvalues of L_k and Ψ the underlying matrix of eigenvectors. Considering $L_k = \Psi^t \Lambda \Psi$, the embedding of a training sample

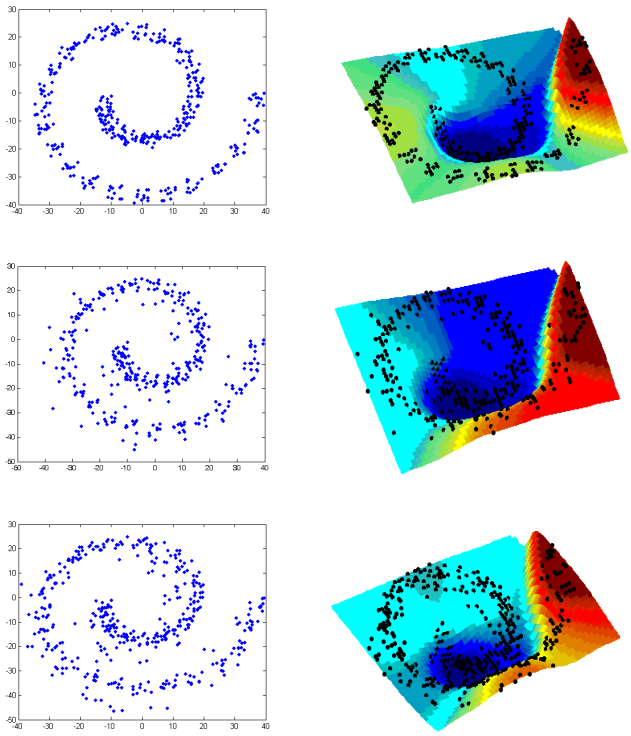

Figure 3. Robustness of the embedding with respect to uniform noise throughout the curvilinear abscissa of the Swiss roll. From top to bottom, the noise is 0%, 15% and 40%.

in $\mathcal{S}$ is $\psi : X_i \mapsto \left(\sqrt{\lambda_1}\,\psi_1(X_i), ..., \sqrt{\lambda_d}\,\psi_d(X_i)\right)'$. d is the intrinsic dimension which corresponds to the largest index $l \in 1, ..., n$ such that $\lambda_l > \delta\lambda_1$ for some $\delta \to 0$ [11]. The diffusion distance can then be expressed in $\mathcal{M}$ as $D_{\mathcal{M}}(X_i, X_j) = \|P_k(i|.) - P_k(j|.)\|^2 = \sum_l \lambda_l \, [\psi_l(X_i) - \psi_l(X_j)]^2$. This distance plays a key role in propagating the labels from the labeled to unlabeled data following the shortest path or the average path (depending on the setting of α).

We define a probabilistic framework which, given a subset of displayed images $\mathcal{D}_1$,...,$\mathcal{D}_t$ until iteration t, makes it possible to explore the manifold $\mathcal{M}$ in order to propose a subset of images $\mathcal{D}_{t+1}$. When we use the unlabeled data by using a transductive algorithm, the heuristics still rely on the following basic assumption: at each iteration, one can select the display in order to refine the current estimate of the decision boundary or one can select the display in order to find uncharted territories in which the actual decision boundary is present. The first display strategy *exploits* our knowledge of the likely position of the decision boundary while the second one *explores* new regions. We believe that any good CBIR system should find the correct balance between exploration and exploitation.

Exploitation: let $\mathcal{D} \subset \mathcal{S}$ and $\mathcal{D}' = \{X \in \mathcal{D}, f_t(X) > 0\}$, (2) is equivalent to :

$$D_{t+1} \leftarrow \arg\max_{\mathcal{D}'} P(\mathcal{D}' \mid \mathcal{D}_t, ..., \mathcal{D}_1) \tag{8}$$

Assuming the data in $\mathcal{D}_{t+1}$ are chosen independently :

$$\begin{aligned} P(\mathcal{D}' \mid \mathcal{D}_t, ..., \mathcal{D}_1) &= \prod_{X_j \in \mathcal{D}'} P(X_j \mid \mathcal{D}_t, ..., \mathcal{D}_1) \\ P(X_j \mid \mathcal{D}_t, ..., \mathcal{D}_1) &\propto \max_{\substack{X_i \in \mathcal{T}_t \\ Y_i = +1}} \frac{1/D_{\mathcal{M}}(X_i, X_j)}{\sum_l 1/D_{\mathcal{M}}(X_i, X_l)}, \end{aligned} \tag{9}$$

Exploration: equivalently, the criteria is similar to (8) but:

$$P(X_j \mid \mathcal{D}_t, ..., \mathcal{D}_1) \propto \min_{\substack{X_i \in \mathcal{T}_t \\ Y_i = +1}} \frac{1/D_{\mathcal{M}}(X_i, X_j)}{\sum_l 1/D_{\mathcal{M}}(X_i, X_l)}, \tag{10}$$

We consider in this work a mixture between the two above strategies where at each iteration t of the interaction process, half of the display (of size 8 in practice) is taken from exploitation and the other set taken from exploration.

4. Nyström Extension

Relevance feedback usually involves databases ranging from many thousands to millions images. The complexity of solving $L_k = \Psi^t \Lambda \Psi$ grows in $O(n^3)$ and on those databases, the problem gets quickly out of hand. For instance, for Corel database ($n = 9.000$) it took about 15 hours to solve the eigenproblem on a standard 64 bits AMD processor of 1.8 GHz, clearly this limits the applicability of the method for large scale databases.

Consider $\mathcal{S}' = \{X_i\}_1^{n'}$ as a subset of $\mathcal{S}$ ($n' \ll n$), n' is chosen such that the above eigenproblem is numerically tractable. The Nyström's extension (see for instance [15]) will then be applied in order to extend the eigen-solution on the whole set $\mathcal{S}$:

$$\psi_l(X) = \sum_{i=1}^{n'} \left(\frac{K(X, X_i)}{\sum_{j=1}^{n} K(X_j, X_i)} \right) \psi_l(X_i), \forall X \in \mathcal{S} \tag{11}$$

Here K is the kernel function used to build the graph Laplacian. In order to show the precision of (11), we randomly select $\mathcal{S}'$ from Corel (see Section 5) with different sizes $n' = 500$, 1.000, 2.000 and 3.000. For a fixed n', we consider 15 different sampling of $\mathcal{S}$, and for each one, we estimate the embedding of $\mathcal{S}'$ using graph Laplacian and we extend on both $\mathcal{S}'$, $\mathcal{S} \setminus \mathcal{S}'$ using the Nyström interpolation. The results reported in Figure (4), show two errors:

1. Curve in green shows: expectations of the interpolation error between (i) graph Laplacian embedding and (ii) Nystroḿ interpolation, both on $\mathcal{S}'$.
2. Curve in blue shows: the same measures but on $\mathcal{S} \setminus \mathcal{S}'$.

In both (1) and (2) the two errors decrease as n' increases and asymptotically converge to the same curve. This clearly corroborates the theoretical statement in [30], which proves that the eigenvector-expansion of the Graph Laplacian converges to the eigenfunctions of the Laplace-Beltrami operator.

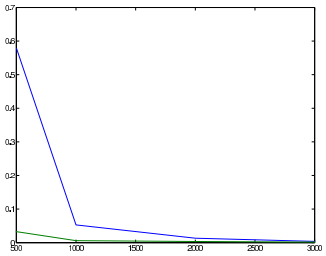

Figure 4. This figure shows the means of the interpolation error using the Nyström's extension. In green: errors on $\mathcal{S}'$. In blue: errors on $\mathcal{S} \setminus \mathcal{S}'$.

5. Performances

In this section, we demonstrate the validity of relevance feedback using our graph Laplacian. We compare it to popular state-of-the-art methods including support vector machines, Bayesian inference and closely related method i.e., graph-cuts. In all these experiments, the size of the display $|\mathcal{D}_t|$ and the number of (simulated) user(s)' responses $|\mathcal{Y}_t|$ are set to 8. The effectiveness is measured as the expected number of images per class which are displayed to the user or equivalently the average number of iterations necessary in order to show a fraction of images per class.

5.1. Databases

Experiments were conducted on simple databases (Olivetti and Swedish) as well as difficult ones (Corel). The Olivetti face database contains 40 persons each one represented by 10 faces. Each face is processed using histogram equalization and encoded using kernel principal component analysis (KPCA) resulting into 20 coefficients. The Swedish set contains 15 categories of leaf silhouettes each one represented by 75 contours. Each contour $\mathcal{C}$ is encoded using 14 coefficients corresponding to the eigenvalues of KPCA on $\mathcal{C}$ [21]. The Corel database contains 90 categories each one represented by 100 images. This database is generic and images range from simple objects to natural scenes with complex background. Each image in this database is encoded simply using a $3D$ RGB color histogram of 125 dimensions. Notice that the classes are very spread so the relevance feedback task is more challenging. For all those databases the ground truth is provided.

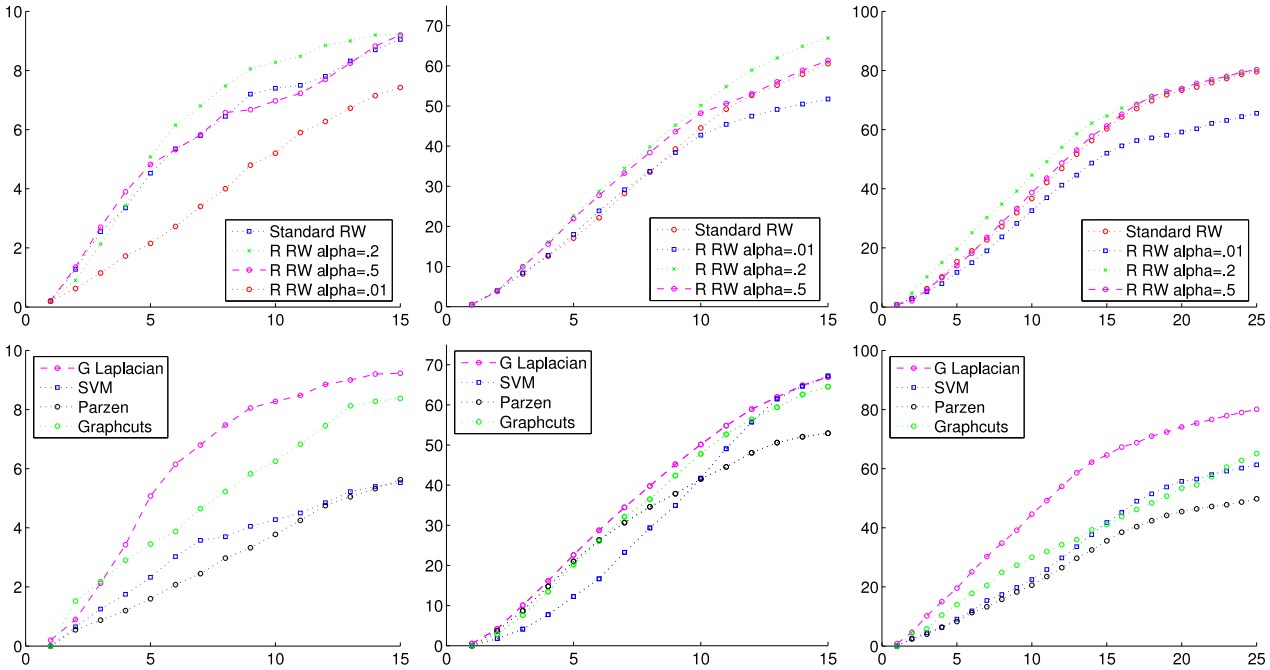

Figure 5. (Top) These figures show the recall for Orl, Swedish and Corel databases for different graph Laplacians. (Bottom) Comparison of Graph Laplacian with respect to SVM, Parzen and Graph-cuts.

5.2. Benchmarking

We evaluate the performance of our RF scheme using the recall. Let Z_t be a random variable standing for the total number of relevant images returned by the CBIR system until iteration t, i.e., those belonging to the user's "class of interest". The recall is defined as $E(Z_t) = \sum_r rP(Z_t = r)$, here the randomness and the expectation of Z_t is taken through different classes of interest. Figures (5, top) show the recall for different graph Laplacians including the standard random walk (RW) and the robust random walk (R RW) for different values of α. The recall reported for the three databases (ORL, Swedish and Corel) show clearly that when $\alpha \ll 1$ (in practice $\alpha = .5$ and $\alpha = .2$), the embedding generated using the graph Laplacian is robust and captures better the topology of the data, and hence the performance follow. Nevertheless, when $\alpha \to 0$ (in practice $\alpha = .01$), the performances degrade as the underlying graph Laplacian implements the most likely path which is more noise-sensitive (see Section 3.2). In all the experiments, the path length k is chosen large enough in order to make the approach robust to noise. In practice and after cross validation we set $k = 10$.

5.3. Comparison

We compared our method to standard representative relevance feedback tools including inductive methods: support vector machines (SVMs), Bayesian inference (based on Parzen windows) and transductive one: Graph cuts. In all these methods, we use the same display strategy (i.e., combined exploration exploitation). We train the SVMs and Parzen classifiers using the triangular kernel as extensive study in [10] showed that SVM based relevance feedback using the triangular kernel achieved far better results than other kernels, so we limit our comparison to SVM and Parzen using this kernel only. Again, for graph Laplacian, the scale parameter of the Gaussian kernel is set as $\sigma = E_{X,X' \in \mathcal{N}_m(X)}\{\|X - X'\|\}$, here $\mathcal{N}_m(X)$ denotes the set of m nearest neighbors of X (in practice $m = 10$). The results reported in Figure (5, bottom), show that in almost all the cases, the recall performances of relevance feedback (using graph-Laplacian) are better than SVMs, Parzen and Graph cuts based RF. Clearly, the use of unlabeled data as a part of transductive learning (in graph Laplacian and graph cuts), makes it possible to improve the performance substantially. Furthermore, the embedding of the data through graph Laplacian makes it possible to capture the topology, so learning the decision rule becomes easier.

6. Conclusion

We introduced in this work an original approach for relevance feedback based on transductive learning using graph Laplacian. This work demonstrates clearly that this semi supervised learning is three-edged sword: it is effective in order (1) to handle transductive learning (in contrast to inductive learning), via the robust s-weighted graph Laplacian which implements the clustering assumption and uses the

unlabeled data as a part of the training process (2) to capture the topology of the data so the similarity measure and the propagation of the labels to unlabeled data is done through the manifold enclosing the data (3) to achieve a clear and consistent improvement with respect to the most powerful and used techniques in relevance feedback including SVMs, Parzen windows and graph cuts. We also demonstrated the efficiency of this approach in order to handle large scale databases using the Nyström extension.

References

[1] Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *PAMI*, 19(7), 1997.

[2] F. Bach. Active learning for misspecified generalized linear models. *Advances in Neural Information Processing Systems (NIPS)*, 19, 2006.

[3] Belkin and Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comp.*, 15(6):1373–1396, 2003.

[4] Belkin and Niyogi. Semi-supervised learning on manifolds. *Machine Learning*, 56:209–239, 2004.

[5] N. Boujemaa, F. Fleuret, V. Gouet, and H. Sahbi. Visual content extraction for automatic semantic annotation of video news. *In the proceedings of the SPIE Conference, San Jose, CA*, 2004.

[6] G. Caenen and E. Pauwels. Logistic regression model for relevance feedback in content based image retrieval. *In proceedings of SPIE*, 4676:49–58, 2002.

[7] Y. Chen, X. Zhou, and T. Huang. One-class svm for learning in image retrieval. *Int'l Conf on Image Processing*, 2001.

[8] I. Cox, M. Miller, T. Minka, and P. Yianilos. An optimized interaction strategy for bayesian relevance feedback. *IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, California*, pages 553–558, 1998.

[9] Y. Fang and D. Geman. Experiments in mental face retrieval. *Proceedings AVBPA 2005, Lecture Notes in Computer Science*, pages 637–646, July 2005.

[10] M. Ferecatu. Image retrieval with active relevance feedback using both visual and keyword-based descriptors. *PhD thesis, Versailles University*, 2005.

[11] M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in euclidean space. In *Proceedings of the 22nd Internatical Conference on Machine Learning (ICML)*, pages 289–296, 2005.

[12] M. Hein, J.-Y. Audibert, and U. von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *http://arxiv.org/abs/math.ST/0608522*, 2006.

[13] Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: query databases through multiple examples. *Int'l Conf. on Very Large Data Bases (VLDB), NY*, 1998.

[14] T. Kurita and T. Kato. Learning of personal visual impression for image database systems. *In the proceedings of the international conference on Document Analysis and Recognition*, 1993.

[15] S. Lafon, Y. Keller, and R. Coifman. Data fusion and multicue data matching by diffusion map. *In IEEE transactions on PAMI*, 2006.

[16] S. MacArthur, C. Brodley, and C. Shyu. Relevance feedback decision trees in content-based image retrieval. *IEEE Workshop CBAIVL*, 2000.

[17] C. Meilhac, M. Mitschke, and C. Nastar. Relevance feedback in surfimage. *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision*, 1998.

[18] H. Narayanan and M. Belkin. On the relation between low density separation, spectral clustering and graph cuts. *In NIPS*, 2006.

[19] R. Picard, T. Minka, and M. Szummer. Modeling user subjectivity in image libraries. *In the proceedings of ICIP*, 1996.

[20] Y. Rui, T. Huang, and S. Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 25–36, 1998.

[21] H. Sahbi. Kernel pca for similarity invariant shape recognition. *In the Journal of Neurocomputing*, 70:3034–3045, 2006.

[22] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. *Proceedings of the ICML*, pages 839–846, 2000.

[23] B. Scholkopf, R. Williamson, A. Smola, J. Taylor, and J. Platt. Support vector method for novelty detection. *Adv. in Neural Information Processing Systems, MIT Press.*, 2000.

[24] M. Seeger. Learning with labeled and unlabeled data. *In Technical Report, University of Edinburgh*, 2001.

[25] A. 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.

[26] D. Spielman and S. Teng. Spectral partioning works: planar graphs and finite element meshes. In *37th Ann. Symp. on Found. of Comp. Science (FOCS)*, pages 96–105. IEEE Comp. Soc. Press., 1996.

[27] K. Tieu and P. Viola. Boosting image retrieval. *IEEE Conf. Computer Vision and Pattern Recognition*, 2000.

[28] S. Tong and E. Chang. Support vector machine active learning for image retrieval. *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, 2001.

[29] N. Vasconcelos and A. Lippman. Learning from user feedback in image retrieval. *in Neural Information Processing Systems MIT press*, 2000.

[30] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 2007.

[31] Y. Zhao, Y. Zhao, and Z. Zhu. Relevance feedback based on query refining and feature database updating in cbir system. *Signal Processing, Pattern Recognition, and Applications*, 2006.

[32] X. Zhou and T. Huang. Relevance feedback in image retrieval: A comprehensive review. *in IEEE CVPR Workshop on Content-based Access of Image and Video Libraries (CBAIVL)*, 2006.