GRAPH LAPLACIAN FOR INTERACTIVE IMAGE RETRIEVAL

Hichem Sahbi[1]*, Patrick Etyngier*[2]*, Jean-Yves Audibert*[2] *and Renaud Keriven*[2]

(1) UMR 5141 CNRS, Telecom ParisTech, France. (2) Certis Lab, ENPC ParisTech, France.

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 a step-by-step labeling of 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 two-folds: it allows us (i) to integrate all the unlabeled images in the decision process and (ii) to robustly capture the topology of the image set. 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.

Index Terms— Statistical Learning, Graph Laplacian and Image retrieval.

1. INTRODUCTION

At least, two interrogation modes are commonly 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. A slight variant is category retrieval which consists in displaying images belonging to the "class of the query". In the second mode (see the pioneering works [1, 2]) the user labels a subset of images as positive and/or negative according to an unknown metric defined in "his mind". Then 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 CBIR system [3, 4]. 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 [5]) and (iii) the self-consistency (and 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*.

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 and the features to the user's feedback. Adapting features might be explicitly achieved or implicitly as a part of the decision rule training. 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 Fig. 1, left). Therefore, further-processing is required in order to extract dimensions with high intrinsic variances. A didactic example, shown in Fig. (1), (the application is searching faces by identity), follows the statement in [6]: *the variance due to the intra-class variability (pose, illumination, etc.) is larger than the inter-class variability (identity)*. Fig. (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[7]. We first model the topology of the image database, including the unlabeled images, using an eigen approximation of the graph Laplacian, then we propagate the labels by projecting the whole dataset

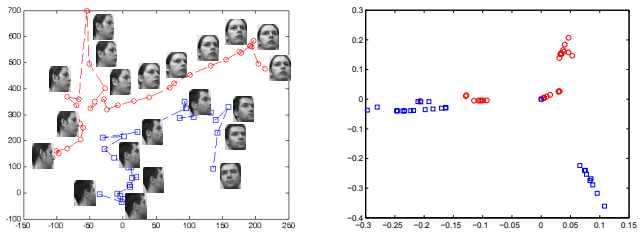

Fig. 1. *(Left) This figure shows the distribution of two classes corresponding to two individuals. It is clear that the intra class variance is larger than the inter class one. (Right) This is the distribution of the same classes inside the manifold trained using graph Laplacian. It is clear that the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

using a linear operator learned on both the labeled and the unlabeled sets. The main contributions of this work are:
(i) 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 [8, 9] (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.

(ii) 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).

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 weighted robust graph Laplacian and the display model. Section 4 provides an experimental study using different databases including specific ones; face databases and also generic databases. We discuss the method and we conclude in Section 5.

2. OVERVIEW OF THE SEARCH PROCESS

Let $\mathcal{S} = \{X_1, ..., X_n\}$, $\{Y_1, ..., Y_n\}$ 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 the 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 $\mathcal{Y}_t \leftarrow \mathcal{L}(\mathcal{D}_t)$.
(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 $\mathcal{S} - \cup_{l=1}^{t} \mathcal{D}_l$ estimating $\text{argmin}_{f:\mathcal{X}\rightarrow\{+1;-1\}} P[f(X) \neq Y]$.
(3) Select the next display $\mathcal{D}_{t+1} \subset \mathcal{S} - \bigcup_{k=1}^{t} \mathcal{D}_k$. 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 $\mathcal{D}_{t+1} \leftarrow \text{argmin}_{\mathcal{D}:\mathcal{D} \bigcap (\bigcup_{l=1}^{t} \mathcal{D}_l) = \emptyset} P[f_{\mathcal{D}}(X) \neq Y]$.

3. GRAPH LAPLACIAN AND RELEVANCE FEEDBACK

Graph Laplacian methods emerged recently as one of the most successful in transductive inference [7], (spectral) clustering and dimensionality reduction. 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 $\mathcal{M}$ denote 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. Transductive Learning using the Graph Laplacian

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. Graph Laplacian is a tranductive method that we hereafter describe. It is based on a neighborhood graph in which the nodes are the input data from both the labeled and unlabeled sets. Let $X_1, \ldots, X_n$ denote these data and let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a symmetrical non-negative function giving the similarity between two input points. The typical kernel is the Gaussian $K(x', x'') = exp(-\|x' - x''\|^2/2\sigma^2)$ and its degree function is defined as $d(x) = \sum_{i=1}^{n} K(X_i, x)$. 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 $K(X_i, X_j)$. 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 $L = D^{-1}W$ 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 $f : \mathcal{X} \rightarrow \mathcal{Y}$, let F be the vector defined as $F_i = f(X_i)$. Now, F is obtained by minimizing $F^t L F$ under the constraints $F_i = Y_i$ for labeled points.

3.2. Our Robust k-step Graph Laplacian

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 Fig. 2, left). In [10], the authors consider instead a graph Laplacian based on the power of L: $L_k = L_{k-1}L$. The matrix L_k 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 (see Fig. 2, middle).

The limitation, mentioned above, motivates the introduction of a new (called robust) graph Laplacian[1], recursively defined as $L_k = [L_{k-1}^{\frac{1}{\alpha}} \times L^{\frac{1}{\alpha}}]^{\alpha}$, $(1/\alpha \in [1, +\infty[)$. Let $L(i,j)^{\frac{1}{\alpha}}$ denote 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[10], 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 over different paths, i.e., $[P_k(j|i)]^{\frac{1}{\alpha}} = \sum_{l=1}^{n} [P_{k-1}(l|i)]^{\frac{1}{\alpha}} \, [P_1(j|l)]^{\frac{1}{\alpha}}$.

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 different random walks. When $\alpha \rightarrow 1$: $P_k(j|i)$ is *the average* transition probability of the k-paths linking X_i to X_j. So L_k implements exactly the one in [10] whereas when $\alpha \rightarrow 0$: $[P_k(j|i)]^{\frac{1}{\alpha}}$ 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. In case $\alpha \in]0,1[$: $[P_k(j|i)]^{\frac{1}{\alpha}}$ 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. Fig. (2, right) shows an example of the application of L_k in embedding of Swiss roll data ($k = 10$ and $\alpha = 0.2$). Clearly, the topology of the data is now preserved.

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

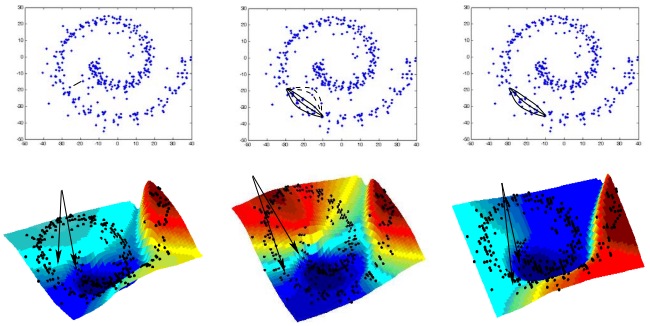

Fig. 2. *The left figures show samples taken from the Swiss roll. (left) A short cut makes the random walk Laplacian embedding very noise sensitive, clearly the variation of the color map does not follow the intrinsic dimension of the actual manifold. (middle) When using the diffusion map, noisy paths 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.*

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 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 $\delta \rightarrow 0$ [10]. 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.

Exploitation: let $\mathcal{D} \subset \mathcal{S}$ and $\mathcal{D}' = \{X \in \mathcal{D}, f_t(X) > 0\}$, the next display is $D_{t+1} \leftarrow \arg\max_{\mathcal{D}'} P(\mathcal{D}' \mid \mathcal{D}_t, ..., \mathcal{D}_1)$. Assuming the data in $\mathcal{D}_{t+1}$ are chosen independently:

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

Exploration: equivalently we replace the $\max$ with $\min$. 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. PERFORMANCE

Experiments were conducted on simple databases : Olivetti ($0.4k$ images) and Swedish ($1,1k$) as well as difficult ones: Corel ($10k$). Each face in Olivetti is encoded using 20 coefficients of KPCA while each contour $\mathcal{C}$ in the Swedish set is encoded using 14 eigenvalues of KPCA on $\mathcal{C}$ [11]. Images in the Corel database are encoded simply using $3D$ RGB color histograms of 125 dimensions so the classes are very spread and the RF task is more challenging.

We evaluate the performance of our RF scheme using the standard recall measure[2]. We compared our method to standard representative RF tools including inductive methods: support vector machines (SVMs), Bayesian inference (based on Parzen windows) and closely related transductive ones: 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 [12] 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 $\sigma = \mathbb{E}_{X,X' \in \mathcal{N}_m(X)}\{\|X - X'\|\}$, here $\mathcal{N}_m(X)$ denotes the set of m nearest neighbors of X (in practice $m = 10$). The results reported in Fig. (3), show that in almost all the cases, the recall performances of RF (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 of the data, so learning the decision rule becomes easier.

5. CONCLUSION

This work introduces an original approach for RF based on transductive learning using graph Laplacian. It demonstrates clearly that the proposed semi-supervised learning method is three-edged sword: it is effective in order (1) to handle transductive learning (in contrast to inductive learning), via the robust 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 and Parzen windows.

[2]This is the fraction of relevant images displayed.

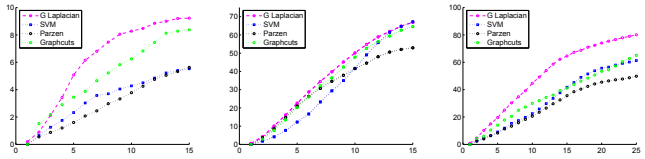

Fig. 3. Comparison, of the recall performance, of Graph Laplacian with respect to SVM and Parzen.

6. REFERENCES

[1] 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.

[2] R.W. Picard, T.P. Minka, and M. Szummer, "Modeling user subjectivity in image libraries," *In the proceedings of the international conference on Image Processing*, 1996.

[3] I.J. Cox, M.L. Miller, T.P. Minka, and P.N. Yianilos, "An optimized interaction strategy for bayesian relevance feedback," *IEEE Conf. on Computer Vision and Pattern Recognition, Santa Barbara*, pp. 553–558, 1998.

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

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

[6] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *PAMI*, vol. 19, no. 7, 1997.

[7] Belkin and Niyogi, "Semi-supervised learning on manifolds," *Machine Learning*, vol. 56, pp. 209–239, 2004.

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

[9] H. Narayanan and M. Belkin, "On the relation between low density separation, spectral clustering and graph cuts," *Advances on Neural information processing systems NIPS*, 2006.

[10] S. Lafon, Y. Keller, and R.R. Coifman, "Data fusion and multi-cue data matching by diffusion map," *to appear in IEEE transactions on Pattern Analysis and Machine Intelligence*, 2006.

[11] H. Sahbi, "Kernel pca for similarity invariant shape recognition," *In the Journal of Neurocomputing*, 2006.

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