Semantic Manifold Learning for Image Retrieval

Yen-Yu Lin[1,2] Tyng-Luh Liu[1] Hwann-Tzong Chen[1,2]

[1]Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[2]Department of CSIE, National Taiwan University, Taipei 106, Taiwan
{yylin, liutyng, pras}@iis.sinica.edu.tw

ABSTRACT

Learning the user's semantics for CBIR involves two different sources of information: the similarity relations entailed by the content-based features, and the relevance relations specified in the feedback. Given that, we propose an *augmented relation embedding* (ARE) to map the image space into a *semantic manifold* that faithfully grasps the user's preferences. Besides ARE, we also look into the issues of selecting a good feature set for improving the retrieval performance. With these two aspects of efforts we have established a system that yields far better results than those previously reported. Overall, our approach can be characterized by three key properties: 1) The framework uses one relational graph to describe the similarity relations, and the other two to encode the relevant/irrelevant relations indicated in the feedback. 2) With the relational graphs so defined, learning a semantic manifold can be transformed into solving a constrained optimization problem, and is reduced to the ARE algorithm accounting for both the representation and the classification points of views. 3) An image representation based on *augmented features* is introduced to couple with the ARE learning. The use of these features is significant in capturing the semantics concerning different scales of image regions. We conclude with experimental results and comparisons to demonstrate the effectiveness of our method.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance feedback*

General Terms

Algorithms, Performance, Experimentation, Management

Keywords

Image Retrieval, Manifold Learning, Dimensionality Reduction, Relevance Feedback, Feature Selection

MM'05, November 6–12, 2005, Singapore.

1. INTRODUCTION

A key ingredient of designing successful Content-Based Image Retrieval (CBIR) systems [4, 16, 31] is how to effectively transform users' interactions into information that could help the underlying retrieval engines to better organize their image data. Different from in information retrieval, the features used in image retrieval are often *visually* characterized, and therefore do not directly connect to the (semantic) concepts implied by users as textual features would do. The semantic gap has been the main challenge to be overcome in CBIR research.

Among the various attempts to deal with the foregoing difficulty, retrieval techniques based on *relevance feedback* [25] are generally considered as a feasible and promising approach, e.g., [9, 10, 12, 13, 23, 24, 28, 29]. Still methods of this kind could differ considerably in the retrieval outcomes. And it brings up two important subjects that would have significant bearings on the query accuracy: 1) the choice of features for representing an image, and 2) the way of capturing the implicit semantic concepts imposed through the few query examples by a user. In this work, we aim to address these two issues by proposing a new manifold-learning scheme with relevance feedback that draws on useful image features to achieve significantly better retrieval performance than that yielded by other existing methods.

1.1 Previous Work

Like in all other classification problems, feature selection when properly done could substantially enhance the retrieval performance. The commonly-indexed features in CBIR are comprehensive, such as shape, texture [9, 10, 24], color, wavelet coefficients [12, 23, 29], and color coherence vectors [19]. The consideration of these features is intuitive, and reasonable for discriminating among images of different categories. They are often implemented as *global* features to describe the respective overall statistics for a *whole* image. Such a practice may cause poor query results when the *region of interest* (ROI) by a user pertains to only a *sub-image*. The situation could further deteriorate when the area of an ROI is relatively small, or a query example has complex background. Alternatively, there are descriptors that are suitable for encoding the local properties of image patches. The SIFT algorithm [15], invented for object recognition and now widely used in vision research, e.g., [8, 14], is one such example, and should be useful in providing additional query cues for a retrieval system.

To take account of relevance feedback, several researchers have explored *supervised learning*. For example, Tong and

Chang [29] propose SVM_{Active} for learning a decision boundary by iteratively adding the most informative (near the boundary) samples as training data. Hoi and Lyu [12] develop a *soft-label* SVM by taking the feedback confidence into consideration in learning the decision boundary. In [28], Tieu and Viola have used AdaBoost to establish a classifier for retrieval, by selecting discriminant features from a very huge candidate pool. Despite these efforts, we note that though SVMs and boosting are effective for classification, the decision boundaries derived by the two schemes would be unstable when the feedback contains only a few image examples for the on-line training. Hence further techniques to address this difficulty are required.

Another possibility is to use the feedback information for adjusting a query vector. Specifically, a query vector can be (dimension-wise) re-weighted [13, 24], moved [13], or expanded [20] to account for users' feedback. Rui et al. [24] propose to iteratively adjust the component weights of a query vector to favor relevant dimensions. Alternatively, in the work of Ishikawa et al. [13], a query can also be modified by considering both the locations and the relevance degree of positive examples. In [20], Porkaew et al. apply clustering to select relevant examples, and then add them into the query representation at each feedback iteration.

The latest trend in CBIR research has been somewhat shifted to recovering the intrinsic structure for a proper image space of reduced dimensionality. Instead of working with the conventional Euclidean space, the main theme is to assume that the images (label and unlabeled) spread as a *manifold*, and the task is to learn the underlying structure. Consequently, a similarity measure can be conveniently computed on the learned manifold. He et al. [10] use geodesic distances to approximate the distances between image pairs along the manifold, and apply Laplacian eigenmaps [2] to preserve such distances. The main drawback is that the mapping is defined only on the set of training data, and thus needs additional mechanisms, such as radial basis function networks, to handle test data. In a related work [9], an incremental learning scheme based on *locality preserving projections* (LPP) [11] has been proposed for semantic relation embedding. Although the mapping derived by LPP is valid for the entire image space, the mapping itself is limited to a linear projection. Furthermore, in both works [9, 10], only one relational graph is used such that the local geometry and feedback relations are not properly represented. As a result, the two schemes may not fully utilize the feedback information in learning the user's semantics.

1.2 Our Approach

Designing an efficient scheme to understand the user's preferences is a nontrivial task in CBIR. We propose an approach that learns a *semantic manifold* by taking account of the multiple aspects of relations among images and the feedback information. In our framework, a similarity relational graph is constructed by exploring the neighborhood of each image, and two feedback relational graphs are created to depict the relevant and irrelevant relations in the feedback. While the similarity property is used as a constraint enforcing the preservation of the local geometry, we make use of the relevant and irrelevant feedback information in a discriminant manner [5], i.e., gathering together relevant pairs and keeping away irrelevant ones after the embedding. In other words, not only the *class* of labeled images but also the *intrinsic structure* of the unlabeled data are considered in learning the semantic manifold. We realize these crucial concepts encoded in the graphs by formulating a constrained optimization problem, and then by solving an equivalent generalized eigenvalue problem. As for indexing images for retrieval, global statistics describe the properties of the whole image and achieve the effectiveness in CBIR, while the local features characterize images by their salient and distinctive regions and are often invariant to certain transformations. Motivated by these observations, we introduce a new image representation to embrace the advantages of the two types of features, and to further improve the retrieval precision by our method.

2. SEMANTIC MANIFOLD LEARNING

The need of dimensionality reduction on analyzing high-dimensional data is unavoidable. Manifold learning is one such technique that aims for finding a constructive way to embed the data from a high-dimensional space into a low-dimensional manifold. Take, for example, the three important works, Isomap [27], LLE [21], and Laplacian eigenmaps [2]. In these methods dimensionality reduction is carried out nonlinearly by investigating the local geometry entailed by the data. Still they all lack an explicit mapping function defined for the entire space, i.e., they cannot directly handle new test data. Bengio et al. [3] have subsequently proposed a new scheme to fix the shortcoming via learning kernel eigenfunctions. Their method can achieve good results, but is too computationally expensive. The LPP by He and Niyogi [11] also shares considerable similarity with Laplacian eigenmaps, except that it has a linear mapping function learned and defined over the whole input space.

All the works discussed above learn data manifolds in an unsupervised manner. While these algorithms are appropriate for data representation and visualization, they do not make the most of the labeled relevance feedback in CBIR. We instead propose a new framework for learning a semantic manifold that best explains the user feedback, comprising only a few labeled examples. Notice that since a structure like this is largely imposed by a user, it implies that the same image database may reside in very different semantic manifolds due to users of diverse preferences.

2.1 Augmented Relation Embedding

To learn a semantic manifold for CBIR, we work on two different sources of information: the similarity relations given by images in the database, and the relevance relations indicated by examples in the feedback. Since the user-provided relevance information can be considered as augmented relations to the data, we choose to use the term, *augmented relation embedding* (ARE), to emphasize this property in our manifold-learning algorithm.

Let $\mathbb{X} \subset \mathbb{R}^n$ be an n-dimensional image feature space, and $\rho : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be some distance function (to be discussed in the next section). A database with m images can then be represented by a data matrix $X = [\mathbf{x}_1\, \mathbf{x}_2\, \cdots\, \mathbf{x}_m] \in \mathbb{R}^{n \times m}$ where $\mathbf{x}_i \in \mathbb{X}$ for $i = 1, \ldots, m$. For the relevance feedback, we use $\mathbf{F}^+$ to denote the set of images returned by the system that are relevant to a query, and $\mathbf{F}^-$ to include the remaining irrelevant images. To characterize the process of ARE, we use three relational graphs (undirected and complete) whose vertices are over the image samples, and a generalized eigenvalue problem, detailed in the following steps.

1. *Construct the similarity relational graph, G^S.* Let the matrix that records the weights over the edges of G^S be $W^S \in \mathbb{R}^{m\times m}$, defined by

$$W^S_{ij} = \begin{cases} e^{-\rho^2(\mathbf{x}_i,\mathbf{x}_j)/t}, & \text{if } \mathbf{x}_i \in k\text{-NN of } \mathbf{x}_j \\ & \text{or } \mathbf{x}_j \in k\text{-NN of } \mathbf{x}_i, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where t is some positive scalar, and k-NN is the abbreviation for the k nearest neighbors.

2. *Construct the feedback relational graphs, G^P and G^N.* The two relational graphs encode pairwise relations in the feedback. In particular, G^P is for the *positively similar* relations, and G^N for the *dissimilar* ones. Their respective weight matrices, $W^P, W^N \in \mathbb{R}^{m\times m}$, can be defined as follows.

$$W^P_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathbf{F}^+ \wedge \mathbf{x}_j \in \mathbf{F}^+, \\ 0, & \text{otherwise;} \end{cases} \quad (2)$$

$$W^N_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathbf{F}^+ \wedge \mathbf{x}_j \in \mathbf{F}^- \\ & \text{or } \mathbf{x}_i \in \mathbf{F}^- \wedge \mathbf{x}_j \in \mathbf{F}^+, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

3. *Embed image space $\mathbb{X}$ into an ℓ-dimensional semantic manifold.* For $\ell \ll n$, find the generalized eigenvectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_\ell$ corresponding to the ℓ largest eigenvalues

$$X[L^N - \gamma L^P]X^T\mathbf{v} = \lambda X L^S X^T \mathbf{v}, \quad (4)$$

where $L^S = D^S - W^S$, and D^S is a diagonal matrix with $D^S_{ii} = \sum_j W^S_{ij}$. Analogously, L^P, D^P, L^N, and D^N can be defined in a similar way. Notice that the scalar γ is added to take care of the possibility of unbalanced feedback. In practice, we have set

$$\gamma \propto \textstyle\sum_{i,j} W^N_{ij} / \sum_{i,j} W^P_{ij}. \quad (5)$$

The parameter γ weighs the importance tradeoff between the positively-similar pairs and the dissimilar ones in the feedback. ($\gamma \geq 1$ is to emphasize the positive information.) Finally, after solving (4) and letting $V = [\mathbf{v}_1\, \mathbf{v}_2 \cdots \mathbf{v}_\ell]$, we have, for each image $\mathbf{x}_i$ in the database, the embedded feature vector $\mathbf{z}_i = V^T\mathbf{x}_i$.

4. *Perform retrieval over the semantic manifold.* While the embedding of the image database is completed, given any arbitrary query image $\bar{\mathbf{x}} \in \mathbb{X}$, we map it onto the manifold by $\bar{\mathbf{z}} = V^T\bar{\mathbf{x}}$. Find the nearest neighbors of $\bar{\mathbf{z}}$ using the Euclidean distance, and those images corresponding to the nearest neighbors will be the top-ranking returns for the query.

We now explain why the steps of ARE can learn a useful semantic manifold for retrieval, and how the generalized eigenvalue problem in (4) guarantees a data embedding that effectively reflects the augmented information. We begin by considering the following optimization problem:

$$\begin{aligned} \text{Maximize} \quad & J(V) = \textstyle\sum_{i,j} ||V^T\mathbf{x}_i - V^T\mathbf{x}_j||^2 (W^N_{ij} - \gamma W^P_{ij}) \\ \text{subject to} \quad & \textstyle\sum_{i,j} ||V^T\mathbf{x}_i - V^T\mathbf{x}_j||^2 W^S_{ij} = 1. \end{aligned} \quad (6)$$

The implication of the above formulation is explicit and reasonable. While the intrinsic structure of the image data is maintained via enforcing the constraint, a feasible V to (6) would project data by reducing the Euclidean distances between each positively-similar pair, and enlarging those between every dissimilar pair. Thus a manifold-learning scheme based on (6) connects the user semantics with the underlying image data in a proper space of reduced dimensionality. We next describe a theorem to complete our discussion on the justification of ARE.

THEOREM 1. *The columns of the optimal $V \in \mathbb{R}^{n\times\ell}$ to the constrained optimization problem (6) are the generalized eigenvectors corresponding to the ℓ largest eigenvalues of (4).*

PROOF. Using the notations in (4) and (6), we have

$$\begin{aligned} J(V) &= \sum_{i,j} ||V^T\mathbf{x}_i - V^T\mathbf{x}_j||^2 (W^N_{ij} - \gamma W^P_{ij}) \\ &= \sum_{i,j} tr\{(V^T\mathbf{x}_i - V^T\mathbf{x}_j)(V^T\mathbf{x}_i - V^T\mathbf{x}_j)^T\}(W^N_{ij} - \gamma W^P_{ij}) \\ &= \sum_{i,j} tr\{V^T(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T V\}(W^N_{ij} - \gamma W^P_{ij}). \end{aligned}$$

Since the trace operator is linear, and $(W^N_{ij} - \gamma W^P_{ij})$ is a scalar, all the terms can be moved inside the trace.

$$\begin{aligned} J(V) &= tr\{V^T \sum_{i,j} (\mathbf{x}_i - \mathbf{x}_j)(W^N_{ij} - \gamma W^P_{ij})(\mathbf{x}_i - \mathbf{x}_j)^T V\} \\ &= 2tr\{V^T[\ (XD^NX^T - XW^NX^T) \\ &\qquad - \gamma(XD^PX^T - XW^PX^T)\ ]V\} \\ &= 2tr\{V^TX(L^N - \gamma L^P)X^TV\}. \end{aligned} \quad (7)$$

After applying a similar analysis to the constraint term, equation (6) can be reformulated as

$$\begin{aligned} \text{Maximize} \quad & J(V) = 2tr\{V^TX(L^N - \gamma L^P)X^TV\} \\ \text{subject to} \quad & 2tr\{V^TXL^SX^TV\} = 1. \end{aligned} \quad (8)$$

Finally, apply the Lagrange multipliers to (8), and set the derivative with respect to V to zero. It follows that the columns of the optimal V are generalized eigenvectors corresponding to the ℓ largest eigenvalues in (4). □

It should be clear now that ARE is a semi-supervised learning technique for dimensionality reduction. The augmented information used in learning a semantic manifold is nicely encoded in the three relational graphs, G^S, G^P, and G^N. Like other manifold-learning methods, the proposed ARE can preserve local geometry by referencing the neighborhood similarity relations in G^S. On the other hand, by exploring the relevance feedback information in G^P and G^N, ARE automatically captures the intrinsic semantics behind the user interactions with a retrieval system.

ARE–Initialization. In general the query-by-example of CBIR starts only with some query image provided by a user. That is, in the inception of manifold learning there should be no feedback information. Consequently, equation (4) is not well defined, and it makes sense to start ARE in an unsupervised manner, i.e., by solving

$$XD^SX^T\mathbf{v} = \lambda XL^SX^T\mathbf{v}. \quad (9)$$

Following the definitions of D^S and L^S, it can be easily verified that with (9), ARE initially behaves like LPP. It then switches to a semi-supervised scheme for learning a sematic manifold when (4) becomes valid.

2.2 Kernel ARE

The query/classification efficiency induced by ARE can sometimes be further improved, especially when the data in the original space are highly nonlinearly distributed. Motivated by the success of *support vector machines* (SVMs) [30], we describe a similar strategy to *kernelize* the linear ARE. The idea is to nonlinearly map the image data to a high-dimensional feature space, and then perform ARE to learn a semantic manifold in that space. Such a generalization is meaningful in the sense that a kernelized ARE would generally achieve better accuracy, and relax the restriction of ARE being only a linear embedding scheme.

Let $\Phi : \mathbb{R}^n \to \mathbb{Y}$ be a nonlinear mapping. Then the image data matrix in the feature space $\mathbb{Y}$ can be denoted as $\Phi(X) \equiv [\Phi(\mathbf{x}_1)\, \Phi(\mathbf{x}_2)\, \cdots\, \Phi(\mathbf{x}_m)]$. Since the analysis mostly involves inner products between pairs of mapped data, it is convenient to work with Mercer kernels instead of worrying about the exact form of Φ. Specifically, we have used the RBF kernel, $\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/c)$ for the experimental results presented in this work.

Consider now a kernel-based optimization problem, the same as (8) except that X is replaced by $\Phi(X)$. Its generalized eigenvalue problem is then given by

$$\Phi(X)[L^N - \gamma L^P]\Phi(X)^T \mathbf{v} = \lambda \Phi(X) L^S \Phi(X)^T \mathbf{v}. \qquad (10)$$

To establish the kernel ARE, we note that the eigenvectors of (10) are in the span of $\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_m)$. In particular, let the eigenvector $\mathbf{v}_i$ of (10) be the ith column of V, and assume the following expansion

$$\mathbf{v}_i = \sum\nolimits_{j=1}^{m} \alpha_{ij} \Phi(\mathbf{x}_j) = \Phi(X)\, \boldsymbol{\alpha}_i, \qquad (11)$$

where $\boldsymbol{\alpha}_i = [\alpha_{i1}\, \alpha_{i2}\, \cdots\, \alpha_{im}]^T$. To this end, it is convenient to define another matrix by $A = [\boldsymbol{\alpha}_1\, \boldsymbol{\alpha}_2\, \cdots\, \boldsymbol{\alpha}_\ell]$, and denote the kernel matrix as $K_{ij} = \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j)$. Furthermore, it can be shown that $V^T \Phi(X) = A^T K$ by element-wise comparison:

$$(V^T \Phi(X))_{ij} = \mathbf{v}_i^T \Phi(\mathbf{x}_j) = (A^T K)_{ij} \qquad (12)$$

for $1 \le i \le \ell$ and $1 \le j \le m$. Therefore the kernelized optimization problem of ARE can be stated as

$$\begin{aligned} \text{Maximize} \quad & U(A) = 2tr\{A^T K (L^N - \gamma L^P) K A\} \qquad (13) \\ \text{subject to} \quad & 2tr\{A^T K L^S K A\} = 1. \end{aligned}$$

The optimal A to (13) would comprise $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, ..., $\boldsymbol{\alpha}_\ell$ that are the generalized eigenvectors corresponding to the ℓ largest eigenvalues of

$$K[L^N - \gamma L^P] K \boldsymbol{\alpha} = \lambda K L^S K \boldsymbol{\alpha}. \qquad (14)$$

Analogously given a query image $\bar{\mathbf{x}}$ for retrieval, the kernel ARE would map the data by $\bar{\mathbf{z}} = V^T \bar{\mathbf{x}}$ with the ith coordinate derived by $\bar{z}_i = \mathbf{v}_i^T \bar{\mathbf{x}} = \sum_{j=1}^{m} \alpha_{ij} \mathbf{k}(\mathbf{x}_j, \bar{\mathbf{x}})$.

3. FEATURES FOR IMAGE RETRIEVAL

Selecting good features is as important as designing an effective learning algorithm for classification problems. In our case, we intend to choose features that are likely to grasp the user's preferences, and general enough for accommodating most retrieval systems. While there is no particular way to categorize image features, we shall divide them into two groups, global and local features. Bear in mind that the main distinction between the two categories of features is not on how they are computed, but on what image scale a feature is set to characterize. We detail in what follows both the global and local features used in our experiments, including their advantages and disadvantages. Then a scheme integrating the two categories is proposed to form *augmented features* for manifold learning with ARE.

3.1 Global Features for CBIR

As we have emphasized, those features used to describe properties concerning a whole image are classified as global. Specifically, in our implementation, we have investigated three types of global features for CBIR.

- *Color.* Features related to color are widely adopted for their simplicity and good performance. We test three kinds of color features: 1) After quantizing the HSV color space, a 64-bin color histogram is evaluated; 2) The first three moments are accordingly extracted from the H, S, and V channels; and 3) Due to the lack of spatial information in the first two, we also add a 128-dimensional *color coherence vector* (CCV) [19] into our global features, to take account of each color's *coherence*.

- *Texture.* Roughly speaking, texture features refer to the image patterns that display homogeneity. They thus play an important role in image indexing of CBIR. In our system, we have considered two kinds of Tamura features, *coarseness* and *directionality*. The former is to measure the distribution about the sizes of image regions with which each pixel is associated, and the latter depicts the information about the magnitudes and the directions of pixel-wise gradients. Similarly, we represent these two features in the form of histograms with 10 and 8 bins, respectively.

- *Wavelet.* Frequency is another aspect of information useful for characterizing images. Among the various techniques, wavelets are deemed to be a powerful tool for capturing both spatial and frequency properties. We apply *discrete wavelet transform* (DWT) to derive a 3-level image decomposition, and then calculate the first two moments of coefficients from the 9 high-frequency sub-bands, i.e., the High/Low, Low/High, and High/High bands in all the three levels.

Having normalized each dimension, we can now represent an image with a 237-dimensional feature vector, computed from the foregoing global descriptors. However, despite the many advantages mentioned above, using global features for CBIR exclusively is not sufficient for ensuring good retrieval performance. In particular, their effectiveness for CBIR could suffer from the following situations.

- When the semantic concepts implied by a user pertain only to sub-images, it is possible the computations of global features may include too many irrelevant factors. As a result, the precision and recall would become worse in that the information used in deriving global features does not fairly reflect the feedback.

- Even for the same semantic concept, the corresponding appearances may differ from image to image, such as poses, scales, pictured viewpoints, or locations of ROI in images. Most global features cannot account for these varieties.

3.2 Local Features for CBIR

Local features are introduced in our implementation to describe properties of size-varying regions associated with *interest points* in an image. It takes two steps to carry out the computations of local features. First, the detection of interest points in a given image is done by using Lowe's *difference-of-Gaussian* (DoG) detector [15], which has been shown to be robust and invariant to scale and rotation. The DoG detector identifies potentially interest points by searching the local extrema in the scale-spaces. After eliminating the unstable ones, the respectively dominating orientation and detected scale are assigned to each of the remaining interest points. Second, we calculate local features from each salient region, identified by the scale, location, and orientation of a detected interest point. Motivated by the good results reported in [18], we consider stacking three different kinds of local features to characterize salient regions.

- *Generalized RGB Color Moments.* The formulation is

$$M_{pq}^{abc} = \int\int_{\Omega} x^p y^q [R(x,y)]^a [G(x,y)]^b [B(x,y)]^c dxdy,$$

  with degree $a+b+c$ and order $p+q$. Let (x,y) denote the relative coordinates with respect to an interest point and its orientation, and Ω be the set of (x,y) within a local region. Setting the degree as 1 or 2, and the order up to 1, we get a 27-dimensional vector.

- *RGB Color histogram.* A color histogram of 64 bins is evaluated to capture the RGB color distribution in the local region of each interest point.

- *SIFT (Scale Invariant Feature Transform) descriptor.* As shown in, e.g., [8, 14, 15, 17], SIFT descriptors are quite effective in representing local image properties. We have used a 128-dimensional SIFT descriptor.

With the above descriptors and proper normalizations, an image would yield a local representation comprising a bag of d local-feature vectors, denoted as $\Gamma = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_d\}$ where d is the number of interest points detected, and $\mathbf{l}_i$ is the local-feature vector for the ith interest point. Since the value of d is image-dependent, the dimensionality of the local representation Γ is not fixed. Thus, for the sake of uniformity that facilitates a similarity measure, we apply the *vector quantization* technique [18, 26] to cluster local-feature vectors resulting from all images into k clusters. In this way, the local representations caused by different numbers of interest points can all be transformed into k-dimensional vectors, where for a given image the value of the ith dimension now records the number of local-feature vectors in Γ being included in the ith cluster.

A proper setting for the value of k is indeed a tradeoff between the degree of precise image representation and the ease of similarity measurement. With a larger k, the differences between two images are more faithfully characterized; however, it becomes inefficient/inappropriate to correlate two images using the *bin-by-bin* similarity measures, e.g., L^2-distance and *Kullback-Leibler divergence*. We instead consider the Earth Mover's Distance (EMD) proposed by Rubner et al. [22], for its nice property in addressing the *cross-bin* dissimilarity. The k-dimensional local representation is therefore converted to the *signature* form used in EMD, where each cluster is represented by its center and the weight (the number of elements in the cluster divided by the total number of elements). Furthermore, the cost between each cluster pair is defined by their geodesic distance, which can be efficiently computed by Floyd's algorithm [6].

To justify the use of EMD with the local representation, we conduct a simple but constructive test by excluding the feedback information and the use of ARE. We begin by preparing a 30-category image set in which each category has 100 images. Those images in the same category are considered relevant, and otherwise, irrelevant. The assumption serves as the ground truth. In the testing of each image, we find its 15-NN (not including the query image) with some similarity measure, and calculate its accuracy. Then the accuracy of each category can be calculated by averaging. The efficiency of using EMD is compared with those yielded by three other distance measures, including the L^2 distance, the negative Bhattacharyya coefficient (BHC), and the dot product (cosine of angle) coupled with *term frequency–inverse document frequency* (TF-IDF) weighting strategy suggested in [26]. Also, the value of k, i.e., the number of feature clusters is set to 3000. The experimental outcomes are shown in Figure 1a. Among the four measures, EMD is clearly the most effective one for our local representation.

3.3 Augmented Features for CBIR

We compare the performance of using either global or local features for CBIR by redoing the experiment in Figure 1a, in which L^2 distance and EMD are respectively employed. The results are shown in Figure 1b (the blue and green curves). Overall the global representation produces better accuracy rates. However, it is worthwhile to note that the two representation schemes complement each other for images of many categories. Taking the most extreme cases into account, e.g., category 20 and 6, we illustrate several images belonging to the two categories in Figures 1c and 1d.

It is evident that the global representation works well for category 20 owing to the consistency in the backgrounds, though the underlying semantic concept is difficult to be identified. On the other hand, using local features achieves good performance for category 6 in that the locally distinctive and unique patterns of jaguars typically appear in small-area sub-images and the backgrounds are also arbitrary and complex. In view of the significantly complementary nature in the query performance, we thus seek a representation that can reasonably include both image properties. Nevertheless, a critical obstacle needed to be surmounted is that the distance measures used for the two types of features are quite different. For example, while the L^2 distance is often used for correlating global feature vectors, it performs poorly for local features. The difficulty prompts the idea of using *augmented features* for CBIR described next.

Given an image, suppose again there are d interest points detected. Therefore its local representation is a bag of local features, $\Gamma = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_d\}$. Let $\mathbf{g}$ be the corresponding global-feature vector for the same image. We define a new representation by augmenting each bag of local features:

$$\mathbf{f}_i = \begin{bmatrix} \omega \mathbf{l}_i \\ (1-\omega)\mathbf{g} \end{bmatrix}, \text{ for } 1 \le i \le d, \tag{15}$$

where ω is a relative weighting factor. Then, the proposed representation, $\widetilde{\Gamma} = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_d\}$, can be handled similarly as for the local representation, including the vector quantization and the use of EMD for distance measurement.

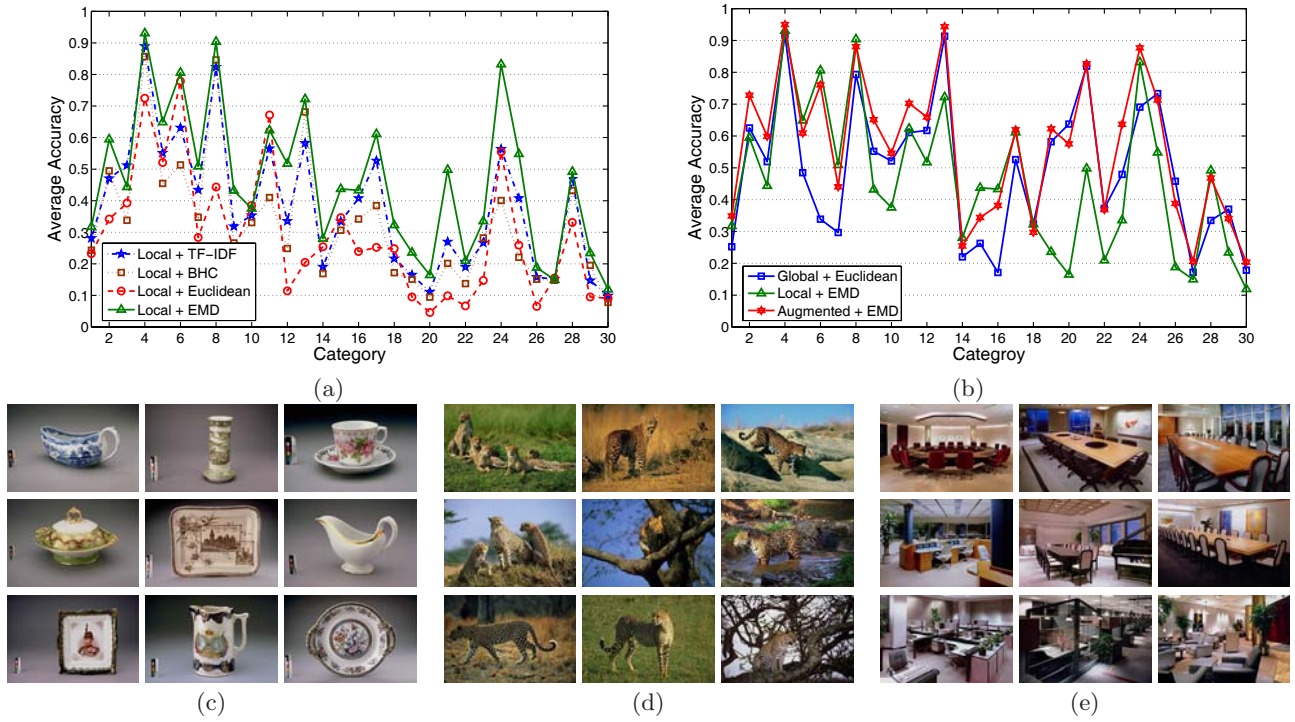

Figure 1: (a) Accuracy comparisons among four similarity measures for local features. (b) Accuracy comparisons for global, local, and augmented features. (c)–(e) Images from category 20, 6, and 23, respectively.

To evaluate the efficiency of the proposed representation, we also carry out the same experiment as in testing the global and the local one. By empirically setting $\omega = 0.5$, the results are shown in Figure 1b (the red curve). It is along the *skyline* of the curves for global and local features. Even for some categories, e.g., 23 in Figure 1e, the augmented features significantly outperform both global and local ones because in these categories the two kinds of statistics, local and global, are meaningful in the similarity measurement.

4. EXPERIMENTS AND DISCUSSIONS

We present several experimental results and comparisons to demonstrate the effectiveness of the proposed manifold-learning algorithm, coupling with the use of augmented features. In Section 4.1 we describe the image dataset used in the experiments. The various settings concerning the performance evaluation metrics, cross validation, and implementation details are given in Section 4.2. We discuss the efficiencies of the three key components in our method, including image representations, learning algorithms, and proper dimensions of embedding spaces for CBIR in Sections 4.3–4.5. Then examples of 2-D visualization for embedding spaces are provided in Section 4.6 to illustrate the progressive improvements through the feedback processes.

4.1 Image Dataset

The COREL dataset is widely used in many CBIR systems, such as [9, 12, 23, 28, 29]. For the sake of evaluations, we also choose the collection for our testing. We empirically select 30 categories of color images, where each consists of 100 samples. Those images in the same category share the same semantic concept, but have their individual varieties. The fact serves as the ground truth in the experiments, i.e., images from the same category are considered relevant, and otherwise irrelevant.

4.2 Evaluation and Implementation Settings

To exhibit the advantages of using our method, we need a reliable way of evaluating the retrieval performance and the comparisons with other systems. We also run cross validation to ensure that the reported results are general and credible. Besides these evaluation settings, different aspects of experimental details are described below.

Evaluation Metrics. Though the *precision-recall* curve is commonly used as a performance measure for retrieval, it is less suitable for the results of CBIR, due to the often relatively low recall [23]. We instead adopt the *precision-scope curve* and the *precision rate* as the performance-evaluation metrics. In this context, the scope specifies the number, N, of top-ranking images returned in response to the user's query, and the precision is the ratio of relevant returns to the scope N. In practice a precision-scope curve records the precision over a range of scopes. The precision rate emphasizes the precision for a particular value of scope, and thus can be viewed as a point on the precision-scope curve. Specifically, we have $N = 20$ for all our experiments. And for those experiments designed for comparing features, we shall use the precision-scope curve to measure the performance, because it gives more comprehensive results. For our other experiments on learning algorithms and dimensionality analysis of the embedding spaces, we prefer the precision rate in that the emphasis should be on the precision differences among

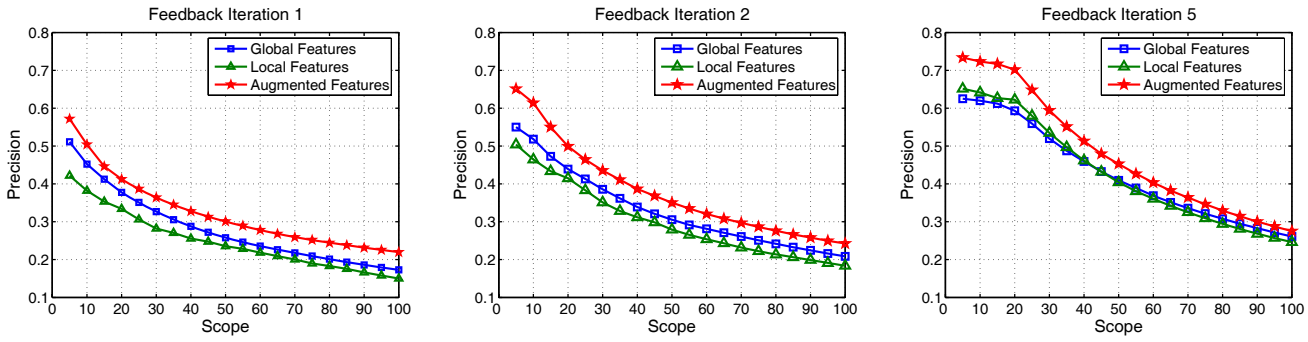

Figure 2: Compare the retrieval performance by applying ARE respectively with the three types of features for image representation: global, local, and augmented features. (a)–(c) Via illustrating with the precision-scope curves, we plot the results in the 1st, 2nd, and 5th feedback iteration, respectively.

all the feedback iterations, and hence a compact description is more appropriate.

Five-Fold Cross Validation. To test our system, we only consider queries that are not in the database. The strategy is practical and meaningful because testing with training data is less persuasive in the evaluation of a learning algorithm. Driven by the query-by-example execution of our system, we use five-fold cross validation to simulate the queries with examples not in the image database. More precisely, we randomly divide all the images into five equal-size sets. In each run of cross validation, we pick one set as the query set, and leave the other four sets as the training data. It implies that the nodes of the relational graphs in the ARE algorithm correspond to images in the training set. The precision-scope curve and precision rate are derived by averaging the results from the five runs of cross validation. We adopt the automatic feedback scheme described in [9] for performance evaluation. For each submitted query, our system retrieves and ranks the images in the training set by iteratively running ARE. At each feedback iteration, the top four relevant and irrelevant images are selected and inserted into the relevant and irrelevant sets, i.e., the $\mathbf{F}^+$ and $\mathbf{F}^-$ in (2) and (3), respectively. Note that the images have been selected in the previous iterations are excluded from later selections. And, with each query, the automatic feedback mechanism is carried out for eight iterations.

Implementation Details. We discuss two issues of implementation details. First, for the concern of numerical stability, we exploit the technique suggested in [1] to avoid singularities encountered in solving the generalized eigenvalue problems. Specifically, we apply PCA to the column space of the data matrix, i.e., X in (4), and keep the 98% information by its low-rank approximation. Second, instead of directly calculating $X[L^N - \gamma L^P]X^T$ in the generalized eigenvalue problem in (4) at each feedback iteration, we compute the result of $\sum_{i,j}(\mathbf{x}_i - \mathbf{x}_j)(W_{ij}^N - \gamma W_{ij}^P)(\mathbf{x}_i - \mathbf{x}_j)^T$, since their equivalence has been shown in (7). In such a way, we take advantage of the sparseness property of W^P and W^N to save the computational resource and prevent the multiplications between large-size matrices. Furthermore, since W^P and W^N are incrementally updated in the feedback iterations, we only take the changed elements in W^P and W^N into account at each iteration.

4.3 Image Features for ARE

In the previous section we have compared the efficiency of using the global, local, and augmented features for retrieval, and reported the results in Figure 1. The experiments are done without using the relevance feedback and any embedding algorithms. Here we again evaluate these three types of image features by testing them with the ARE algorithm. Via illustrating with the precision-scope curves, their performance in the 1st, 2nd, and 5th feedback iteration is plotted in Figures 2a–2c, respectively. Based on the results shown in the diagrams, we observe: 1) The augmented features are more efficient than the other two classes in all the iterations; 2) Owing to the increasing number of feedback images in the latter iterations, the precision is improved over the entire range of the scope; and 3) In the latter feedback iterations, the precision decays slightly within the small-scale scope, as shown in Figure 2c. The phenomenon may be caused by a better fitting of ARE with more feedback information.

4.4 Manifold Learning Schemes

To demonstrate the power of the proposed ARE algorithm in learning the semantic concepts from feedback examples, we compare its retrieval performance with that of a related scheme, namely, the incremental Locality Preserving Projections [9]. Both the two algorithms measure similarities locally based on the manifold assumption, and are designed for learning the semantic space via solving eigenvalue problems. The critical difference between the two schemes is that the incremental LPP maintains only a graph for recording both the neighborhood and feedback relations at the same time, while the ARE treats the neighborhood relation as a constraint and formulates, in a discriminant manner, feedback information into an objective function.

Besides incremental LPP and ARE, the kernel ARE is also included in the comparisons. Together with the three possible choices of feature representations, we conduct nine experiments about the precision of learning a semantic manifold. By iteratively adding the user's feedback, the corresponding precision results of the three learning schemes are respectively shown in Figures 3a–3c, ordered by the image features used. We next highlight some remarks on the experimental results in Figures 3.

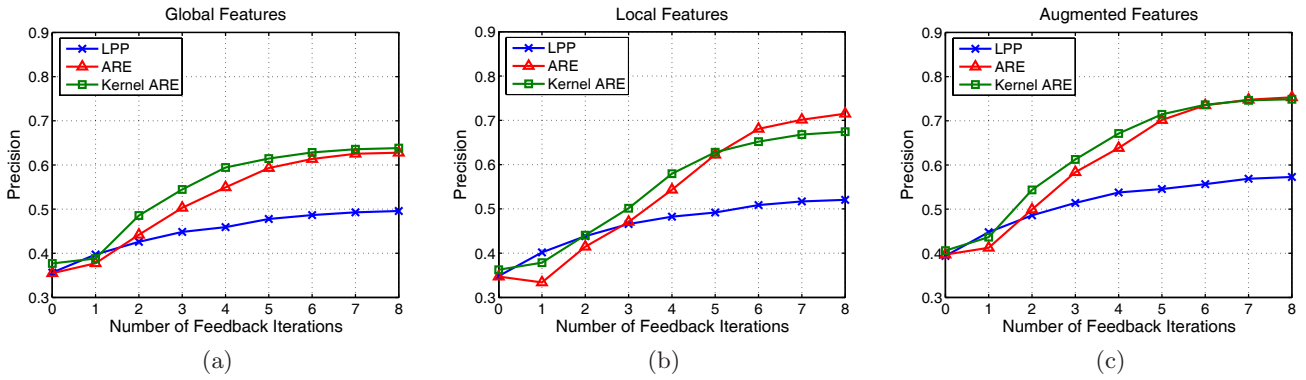

Figure 3: Performance evaluation of the three learning algorithms in learning the semantic concepts from the feedback, by using (a) global, (b) local, and (c) the augmented features.

- No matter what kind of image representation is used, the ARE and kernel ARE significantly outperform the incremental LPP especially in the latter feedback iterations. The incremental LPP is formulated based on only one neighborhood graph. For a node in the graph that corresponds to a labeled relevant image, it cannot differentiate other labeled relevant images from its neighbors (can be either relevant or irrelevant). Meanwhile, it also cannot distinguish other labeled irrelevant images from those which are not its neighbors (again can be either relevant or irrelevant). Thus, incremental LPP is not well account for users' feedback. On the contrary, ARE uses two additional graphs to encode the augmented relations from the feedback, and effectively transforms these relations into a constrained optimization problem. The underlying semantics by a user are therefore faithfully retained.

- Kernel ARE in most cases outperforms ARE except for the latter iterations of using local features (though they eventually converge to a similar degree of precision). The efficiency boost is owing to the fact that kernel ARE is a nonlinear scheme, and performs manifold learning over a high-dimensional feature space. However, the main drawback of kernel ARE is that it takes considerably longer time than ARE in learning the embedding. Especially, the value of variance used in the RBF kernel function is often determined by time-consuming brute force searching. In addition, for a testing image, its inner products with all the training images in the high-dimensional space need to be computed to find its coordinates in the embedding space.

4.5 Embedding Dimensions

The dimensionality of the embedding space is critical to the retrieval precision and the time-complexity efficiency for the search of nearest neighbors in our system. We argue that, for a good manifold-learning algorithm, the following two requirements for the learned manifold are essential and favorable to be satisfied. First, the precision should converge rapidly with respect to the increasing of the dimension. This property ensures the correctness of a system, the compactness of image representations, and the efficiency of similarity search in the low-dimensional embedding space. Second, it is useful to have a broad range of dimensions that is *optimal* for system precision. Furthermore, the change in the precision along the dimension axis should be stable and smooth. The optimal dimension for embedding can therefore be conveniently spotted. Also, for image retrieval with relevance feedback, it is preferable the optimal ranges of embedding dimensions are mostly overlapped for all the feedback iterations. Thus a common value of dimensionality can be applied to each iteration.

To verify whether our method has the above properties for an efficient manifold-learning scheme, we respectively evaluate the precisions of the ARE in different feedback iterations over a range of embedding dimensions. The results are shown in Figure 4. Besides the iterative improvement and convergency in precision over the feedback iterations, we also observe that ARE satisfies the requirements: the precision converges (along the curve) near the dimensions of $30 \sim 40$, and a broad optimal region around dimensions $40 \sim 100$ is shared by all the feedback iterations.

4.6 Visualization of Semantics

To gain insight into ARE, we display the learned semantic manifold in a 2-D plane. However, ARE does not perform well to embed the manifold into such a low-dimensional space (see Figure 4). Instead of directly embedding into a 2-D space for visualization, we use ARE to embed the semantic manifold into a 30-dimensional space, and project the points on a plane via *multidimensional scaling* (MDS) [7], which preserves the inter-point distances of the 30-D space as faithfully as possible. In Figures 5a and 5b, we show the two queries with the respective semantic concepts of *firework* and *office interiors*, and their learned semantic manifolds. The images of the two queries are depicted in the first row. The second row includes the initial embedding spaces (without any feedback). In the last two rows, the semantic manifolds learned after the 3rd and the 8th feedback iterations are given. In each figure, the red points represent the images relevant to the query, and the green points stand for the irrelevant ones. The four magenta and cyan points respectively denote the relevant and irrelevant feedback that will be returned to the system. The region centered at the query point is zoomed-in to reveal the detail of 100-NN of the query. Note that the relevant (red) points progressively gather together while the irrelevant points keep away from the relevant ones, especially in the region around the query.

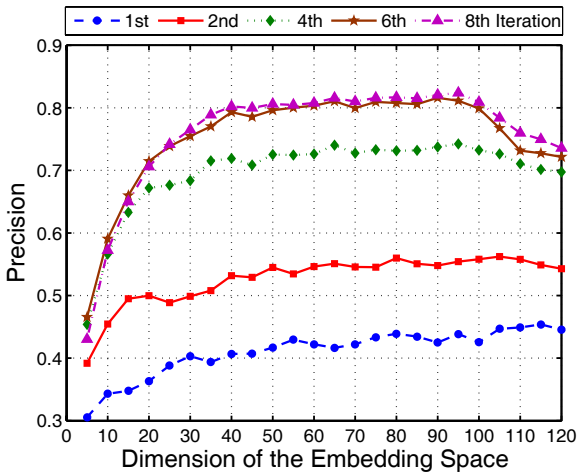

Figure 4: The dimensionality effect of ARE in learning the semantic manifold. We depict the precisions in different feedback iterations with the dimensionality from 5 to 120 of the embedding space. Note that, using 40 to 100 dimensions, ARE uniformly gives stable and reliable results for all the feedback iterations.

Besides the quantitative results, the illustrations of the embedding spaces also demonstrate the effectiveness of the proposed ARE in learning the semantic manifolds.

5. CONCLUSION

We have presented a framework for learning a semantic manifold of CBIR, and applied the technique to capture the user's preferences from few feedback examples. Our method is further consolidated with the use of augmented features, designed to more precisely characterize an image by both its global and local properties. The ARE completes the embedding in a transductive manner by taking both the class of the labeled images and the intrinsic structure of the unlabeled ones into account. The promising experimental results and several useful comparisons justify their use in CBIR with relevance feedback. Owing to the generality of ARE, we consider to connect the algorithm to process other multimedia data such as audio and video for our future work.

6. ACKNOWLEDGMENTS

This work was supported by grants 93-2213-E-001-018, 94-2213-E-001-005 and 94-EC-17-A-02-S1-032.

7. REFERENCES

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenface vs. fisherfaces: Recognition using class-specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Neural Information Processing Systems*, 2001.

[3] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *Neural Information Processing Systems*, 2003.

[4] C. Carson, M. Thomas, S. Belongie, J. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *Visual Information Systems*, pages 509–516, 1999.

[5] H.-T. Chen, H.-W. Chang, and T.-L. Liu. Local discriminant embedding and its variants. In *Int'l Conference on Computer Vision and Pattern Recognition*, pages II: 846–853, 2005.

[6] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2nd edition, 2001.

[7] T. Cox and M. Cox. *Multidimentional Scaling*. Chapman & Hall, London, 1994.

[8] K. Grauman and T. Darrell. Efficient image matching with distributions of local invariant features. In *Int'l Conference on Computer Vision and Pattern Recognition*, pages II: 627–634, 2005.

[9] X. He. Incremental semi-supervised subspace learning for image retrieval. In *ACM Conference on Multimedia*, pages 2–8, 2004.

[10] X. He, W.-Y. Ma, and H.-J. Zhang. Learning an image manifold for retrieval. In *ACM Conference on Multimedia*, pages 17–23, 2004.

[11] X. He and P. Niyogi. Locality preserving projections. In *Neural Information Processing Systems*, 2003.

[12] C.-H. Hoi and M. Lyu. A novel log-based relevance feedback technique in content-based image retrieval. In *ACM Conference on Multimedia*, pages 24–31, 2004.

[13] Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples. In *International Conference on Very Large Data Bases*, pages 218–227, 1998.

[14] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM Conference on Multimedia*, pages 869–876, 2004.

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l Journal of Computer Vision*, 60(2):91–110, 2004.

[16] W.-Y. Ma and B. Manjunath. Netra: A toolbox for navigating large image databases. In *Multimedia Systems*, volume 7, pages 184–198, 1999.

[17] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Int'l Conference on Computer Vision and Pattern Recognition*, pages 275–263, 2003.

[18] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Euro. Conference on Computer Vision*, pages 71–84, 2004.

[19] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *ACM Conference on Multimedia*, pages 65–73, 1996.

[20] K. Porkaew and K. Chakrabarti. Query refinement for multimedia similarity retrieval in MARS. In *ACM Conference on Multimedia*, pages 235–238, 1999.

[21] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

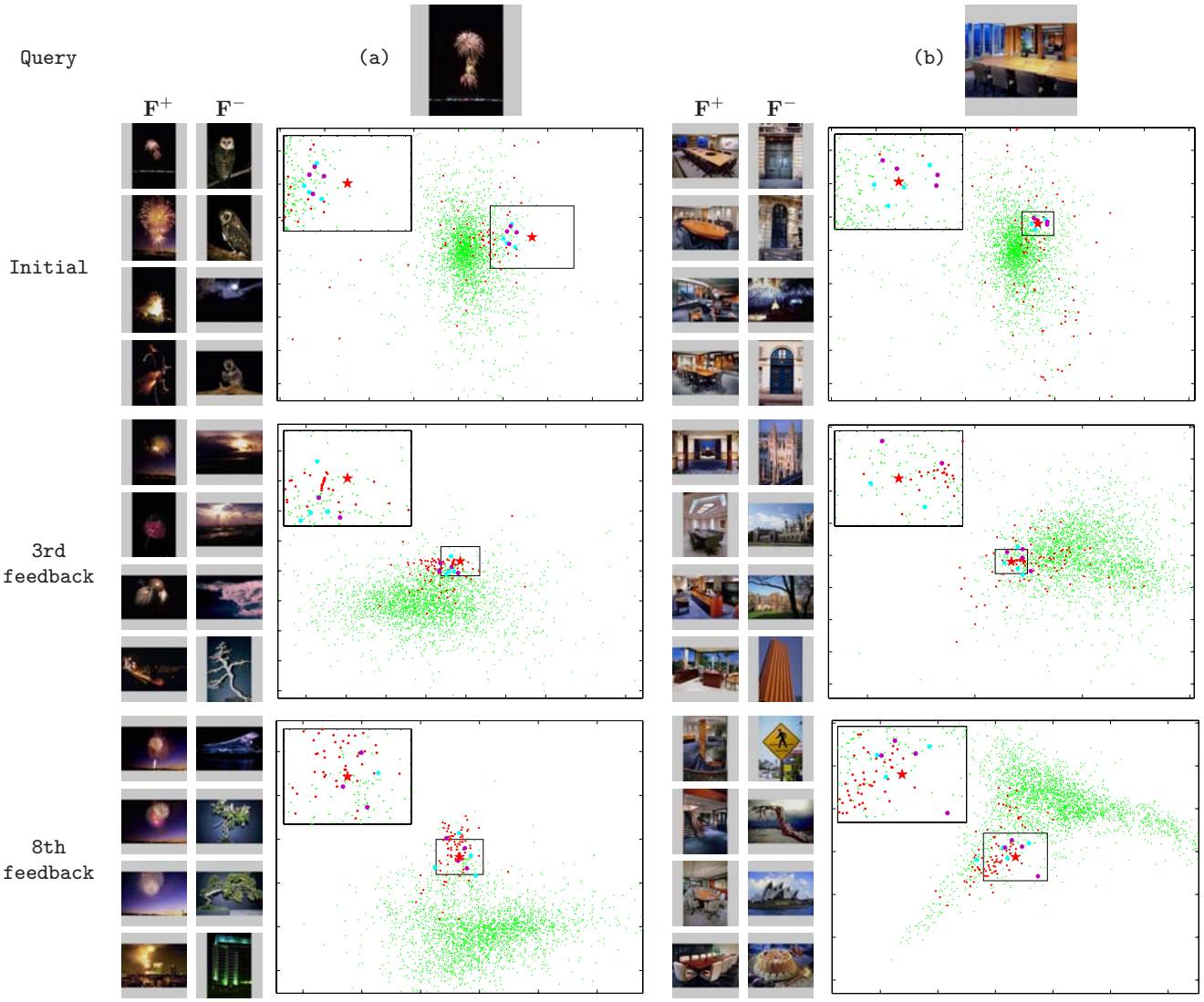

Figure 5: The learned semantic manifolds for the queries (a) *firework* & (b) *office interiors*. Each (zoomed-in) rectangle encloses the 100-NN of the query. Through the iterative embedding, the relevant images (red points) progressively gather around the query while the irrelevant ones (green points) are kept far away.

[22] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *Int'l Journal of Computer Vision*, 40(2):99–121, 2000.

[23] Y. Rui and T. Huang. Optimizing learning in image retrieval. In *Int'l Conference on Computer Vision and Pattern Recognition*, pages 236–243, 2000.

[24] Y. Rui, T. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in mars. In *Int'l Conference on Image Processing*, pages 815–818, 1997.

[25] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Company, 1982.

[26] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Int'l Conference on Computer Vision*, pages 1470–1477, 2003.

[27] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[28] K. Tieu and P. Viola. Boosting image retrieval. In *Int'l Conference on Computer Vision and Pattern Recognition*, pages 1228–1235, 2000.

[29] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM Conference on Multimedia*, pages 107–118, 2001.

[30] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[31] J. Wang, J. Liu, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(9):847–963, 2001.