Learning an Image Manifold for Retrieval

Xiaofei He*,   Wei-Ying Ma, and Hong-Jiang Zhang

Microsoft Research Asia
Beijing, China, 100080
{wyma,hjzhang}@microsoft.com

*Department of Computer Science,
The University of Chicago
xiaofei@cs.uchicago.edu

ABSTRACT

We consider the problem of learning a mapping function from low-level feature space to high-level semantic space. Under the assumption that the data lie on a submanifold embedded in a high dimensional Euclidean space, we propose a relevance feedback scheme which is naturally conducted only on the image manifold in question rather than the total ambient space. While images are typically represented by feature vectors in $\mathbf{R}^n$, the natural distance is often different from the distance induced by the ambient space $\mathbf{R}^n$. The geodesic distances on manifold are used to measure the similarities between images. However, when the number of data points is small, it is hard to discover the intrinsic manifold structure. Based on user interactions in a relevance feedback driven query-by-example system, the intrinsic similarities between images can be accurately estimated. We then develop an algorithmic framework to approximate the optimal mapping function by a Radial Basis Function (RBF) neural network. The semantics of a new image can be inferred by the RBF neural network. Experimental results show that our approach is effective in improving the performance of content-based image retrieval systems.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Algorithms, Indexing methods.*

General Terms

Algorithms, Management, Performance, Experimentation.

Keywords

Image Retrieval, Semantic Space, Manifold Learning, Dimensionality Reduction, Riemannian Structure

1. INTRODUCTION

Content-Based Image Retrieval (CBIR) [3][9][12][14][21] is a long standing research problem in computer vision and information retrieval. Most of previous image retrieval techniques build on the assumption that the image space is Euclidean. However, in many cases, the image space might be a non-linear sub-manifold which is embedded in the ambient space. Intrinsically, there are two fundamental problems in image retrieval: 1) How do we rep-

resent an image? 2) How do we judge similarity?

One possible solution to these two problems is to learn a mapping function from the low-level feature space to the high-level semantic space. The former is not always consistent with human perception while the latter is what image retrieval system desires to have. Specifically, if two images are semantically similar, then they are close to each other in semantic space. In this paper, our approach is to recover semantic structures hidden in the image feature space such as color, texture, etc.

In recent years, much has been written about relevance feedback in content-based image retrieval from the perspective of machine learning [16][17][18][19][20], yet most learning methods only take into account current query session and the knowledge obtained from the past user interactions with the system is forgotten. To compare the effects of different learning techniques, a useful distinction can be made between short-term learning within a single query session and long-term learning over the course of many query sessions [6]. Both short- and long-term learning processes are necessary for an image retrieval system though the former has been the primary focus of research so far. We present a long-term learning method which learns a radial basis function neural net-work for mapping the low-level image features to high-level semantic features, based on user interactions in a relevance feedback driven query-by-example system.

As we point out, the choice of the similarity measure is a deep question that lies at the core of image retrieval. In recent years, manifold learning [1][4][11][13][15] has received lots of attention and been applied to face recognition [7], graphics [10], document representation [5], etc. These research efforts show that manifold structure is more powerful than Euclidean structure for data representation, even though there is no convincing evidence that such manifold structure is accurately present. Based on the assumption that the images reside on a low-dimensional submanifold, a geometrically motivated relevance feedback scheme is proposed for image ranking, which is naturally conducted only on the image manifold in question rather than the total ambient space.

It is worthwhile to highlight several aspects of the framework of analysis presented here:

(1) Throughout this paper, we denote by image space the set of all the images. Different from most of previous geometry-based works which assume that the image space is a Euclidean space [8][12], in this paper, we make a much weaker assumption that the image space is a Riemannian manifold em-

bedded in the feature space. Particularly, we call it image manifold. Generally, the image manifold has a lower dimensionality than the feature space. The metric structure of the image manifold is induced but different from the metric structure of the feature space. Thus, a new algorithm for image retrieval which takes into account the intrinsic metric structure of the image manifold is needed.

(2) Given enough images, it is possible to recover the image manifold. However, if the number of images is too small, then any algorithm can hardly discover the intrinsic metric structure of the image manifold. Fortunately, in image retrieval, we can make use of user provided information to learn a semantic space that is locally isometric to the image manifold. This semantic space is Euclidean and hence the geodesic distances on the image manifold can be approximated by the Euclidean distances in this semantic space. This intuition will be strengthened in our experiments.

(3) There are two key algorithms in this framework. One is the retrieval algorithm on image manifold, and the other is an algorithm for learning a mapping function from feature space (color, texture, etc.) to high-level semantic space. The learning algorithm will gradually "flat" the image manifold, and make it better consistent with human perception. That is, if two images are close (in the sense of Euclidean metric) to each other, they are semantically similar to each other. Here, by "flat" we mean that the image manifold will be ultimately equipped with a flat Riemannian metric defined on it, at which time we call it semantic space.

The rest of this paper is organized as follows: Section 2 describes the proposed retrieval algorithm on image manifold. Section 3 describes the proposed framework for learning a semantic space to represent the underlying image manifold. The experimental results are shown in Section 4. We give concluding remarks in Section 5.

2. RELEVANCE FEEDBACK ON IMAGE MANIFOLD

In many cases, images may be visualized as points drawn on a low-dimensional manifold embedded in a high-dimensional Euclidean space. In this paper, our objective is to discover the image manifold by a locality-preserving mapping for image retrieval. We propose a geometrically motivated relevance feedback scheme for image ranking, which is conducted on the image manifold, rather than the total ambient space.

2.1 The Algorithm

Let Ω denote the image database and R denote the set of query images and relevant images provided by the user. Our algorithm can be described as follows:

1. **Candidate generation**. For each image $\mathbf{x}_i \in R$, we find its k-nearest neighbors $C_i = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k\}$, $\mathbf{y}_j \in \Omega$ (those images in R are excluded from selection). Let $C = C_1 \cup C_2 \cup \ldots \cup C_{|R|}$. We call C candidate image set. Note that $R \cap C = \varnothing$.

2. **Construct subgraph**. Construct a graph $G(V)$, where $V = R \cup C$. The distance between any two images $\mathbf{x}_i, \mathbf{x}_j \in V$ is measured as follows:

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|_2 & if \; \left\| \mathbf{x}_i - \mathbf{x}_j \right\|_2 < \varepsilon \\ \infty & otherwise \end{cases}$$

where ε is a suitable constant. The choice of ε reflects our definition of *locality*. We put an edge between $\mathbf{x}_i$ and $\mathbf{x}_j$ if $dist(\mathbf{x}_i, \mathbf{x}_j) \neq \infty$. Since the images in R are supposed to have some common semantics, we set their distances to zero. That is, $dist(\mathbf{x}_i, \mathbf{x}_j) = 0$, $\forall \mathbf{x}_i, \mathbf{x}_j \in R$. The constructed graph models the local geometrical structure of the image manifold.

3. **Distance measure on image manifold**. To model the geodesic distances between all pairs of image points on the image manifold M, we find the shortest-path distances in the graph G. The length of a path in G is defined to be the sum of link weights along that path. We then compute the geodesic distance $dist_G(\mathbf{x}_i, \mathbf{x}_j)$ (i.e. the shortest path length) between all pairs of vertices of i and j in G, using Floyd's $O(|V|^3)$ algorithm.

4. **Retrieval based on geodesic distance**. To retrieve the images most similar to the query, we simply sort them according to their geodesic distances to the query. The top N images are presented to the user.

5. **Update query example set**. Add the relevant images provided by the user into R. Go back to step 1 until the user is satisfied.

2.2 Geometrical Justification

Our algorithm deals with finite data sets of points in $\mathbf{R}^n$ which are assumed to lie on a smooth submanifold M with low dimensionality. The algorithm attempts to recover M given only the data points. A crucial stage in the algorithm involves estimating the unknown geodesic distance in M between data points in terms of the graph distance with respect to some graph G constructed on the data points.

The natural Riemannian structure on M (induced from the Euclidean metric on $\mathbf{R}^n$) gives rise to a manifold metric d_M defined by:

$$d_M(\mathbf{x}, \mathbf{y}) = \inf_r \{length(r)\}$$

where r varies over the set of (piecewise) smooth arcs connecting $\mathbf{x}$ to $\mathbf{y}$ in M. Note that $d_M(\mathbf{x}, \mathbf{y})$ is generally different from the Euclidean distance $\|\mathbf{x}-\mathbf{y}\|$. Our algorithm makes use of a graph G on the data points. Given such a graph we can define a metric, just on the set of data points. Let $\mathbf{x}, \mathbf{y}$ belong to the set $\{\mathbf{x}_i\}$. We define:

$$d_G(\mathbf{x}, \mathbf{y}) = \textit{the length of the shortest path between } \mathbf{x} \textit{ and } \mathbf{y}$$

Given the data points and graph G, one can compute d_G without knowledge of the manifold M. Bernstein *et al.* [2] show that the two distance metrics (d_M and d_G) approximate each other arbitrarily closely, as the density of data points tends to infinity.

Here, we give a simple example to show the advantage of geodesic distances on manifold over Euclidean distance, and the advantage of semantic space over low-level image feature space. Figure 1 shows a spiral on a plane. Consider that the images of our concern are sampled from the spiral. Clearly, it is a one-dimensional manifold. Figure 1(a) shows the Euclidean distance between data points A and B. Figure 1(b) shows the geodesic distance along the spiral. In this example, the intrinsic geometrical structure can only be characterized by the geodesic distance.

In many real world applications, one is often confronted with the problem that the number of sample points is too small to describe the underlying topology of the data. In this case, the geodesic
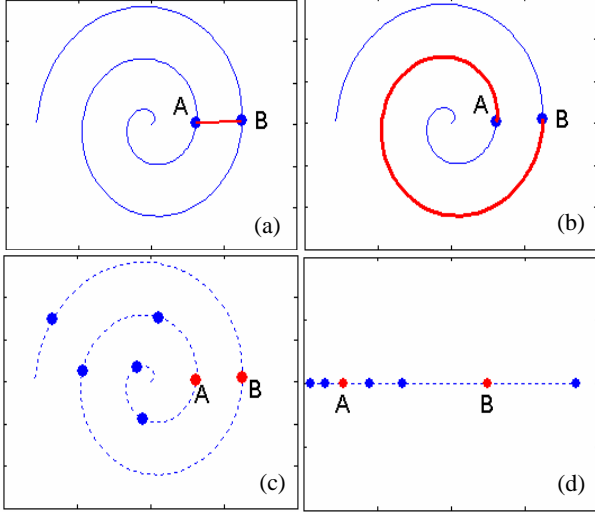
Figure 1. (a) Euclidean distance between data points _A_ and _B_. (b) Geodesic distance between data points _A_ and _B_. (c) Seven data points sampled from the spiral. The geodesic distances between them can not be accurately estimated. (d) 1-D representation of the spiral.

distance on the image manifold can not be accurately estimated, as can be seen from Figure 1(c). Fortunately, in image retrieval, the user provided information can help us recover the underlying structure of the image manifold. In the next section, we will describe how to learn a continuous function which maps the data points (images) into a semantic space in which the Euclidean distances between images are consistent with human perception, as illustrated in Figure 1(d). The _nonlinear_ Riemannian structure of the manifold (Figure 1(c)) can be inferred from the _linear_ Euclidean structure of the semantic space (Figure 1(d)).

It might be more interesting to consider this example in image retrieval domain. Suppose the point _A_ is the query image and the other six points denote the images in database. If we conduct the retrieval in the low-level feature space (Figure 1(c)), the point _B_ will be selected as relevant image, no matter what distance metric we use, Euclidean or geodesic. This is because the intrinsic Riemannian structure of the image manifold can not be accurately detected due to the lack of sufficient sample points. However, if the retrieval is conducted in the semantic space (Figure 1(d)), the point _B_ will never be selected as relevant image. This is because that, by incorporating user provided information, the intrinsic Riemannian structure of the image manifold can be accurately detected. Clearly, the retrieval in semantic space is more consistent with human perception.

3. USING MANIFOLD STRUCTURE FOR IMAGE REPRESENTATION

In the previous section, we have described an algorithm to retrieve the user desired images by modeling the underlying geometrical structure of the image manifold. One problem of this algorithm is that, if the number of sample images is very small, then it is difficult to recover the image manifold. In this case, we propose a long-term learning approach to discover the true topology of the image manifold using user interactions. To be specific, we aim at

mapping each image into a semantic space in which the distances between the images are consistent with human perception.

The problem we are going to solve can be simply stated below:

> Let _S_ denote the low-level feature space, and _T_ denote the semantic space. Learn a nonlinear mapping function from _S_ to _T_,
>
> $$f : \mathbf{x} \rightarrow \mathbf{z} \qquad (\mathbf{x} \in S, \mathbf{z} \in T)$$
>
> which preserves the local Riemannian structure of the low-level feature space.

Our proposed solution consists of three steps:

1. Inferring a semantic matrix $B_{m \times m}$ from user interactions, whose entries are the distances between pairs of images in semantic space _T_. _m_ is the number of images in database.

2. Find _m_ points $\{\mathbf{z}_1, \mathbf{z}_2,\ldots, \mathbf{z}_m\} \subset \mathbf{R}^k$ which preserve pairwise distances specified in $B_{m \times m}$. Laplacian eigenmaps [1] is used to find such an embedding. The space in which the _m_ points $\{\mathbf{z}_1, \mathbf{z}_2,\ldots, \mathbf{z}_m\}$ are embedded is called _LE semantic space_ in the rest of the paper. The user provided information is incorporated into the LE semantic space. Note that, the LE semantic space is only defined on the image database. In other words, for a new image outside the database, it is unclear how to evaluate its coordinates in the LE semantic space.

3. Given _m_ pair vectors, $(\mathbf{x}_i , \mathbf{z}_i)$ $(i = 1,2,\ldots,m)$, where $\mathbf{x}_i$ is the image representation in low-level feature space, and $\mathbf{z}_i$ is the image representation in LE semantic space, train a radial basis function (RBF) neural network _f_ that accurately predicts future **z** value given **x**. Hence $f(\mathbf{x})$ is a semantic representation of **x**. The space obtained by _f_ is called _RBFNN semantic space_. Note that, $f(\mathbf{x}_i) \approx \mathbf{z}_i$. That is, RBFNN semantic space is an approximation of the LE semantic space. However, RBFNN semantic space is defined everywhere. That is, for any image (either inside or outside the database), its semantic representation can be obtained from the mapping function.

We describe the detail of these steps in the following.

3.1 Inferring a Distance Matrix in Semantic Space from User Interactions

In this section, we describe how to infer a distance matrix in semantic space from user interactions. Some previous work could be found in [6]. Here, we present a simple method to update the distance matrix gradually.

Let _B_ denote the distance matrix, $B_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$. Intuitively, the images marked by the user as positive examples in a query session share some common semantics. Therefore, we can shorten the distances between them. Let _S_ denote the set of positive examples, $S = \{s_1, s_2, \ldots, s_k\}$. We can adjust the distance matrix as follows:

$$B_{s_i s_j} \leftarrow B_{s_i s_j} / \alpha \quad (s_i, s_j \in S)$$

where α is a suitable constant greater than 1. Similarly, we can lengthen the distances between the positive examples and negative examples, as follows:

$$R_{s_i t_j} \leftarrow R_{s_i t_j} \times \beta \quad (s_i \in S, t_j \in T)$$

where $T = \{t_1, t_2, \ldots, t_k\}$ is the set of negative examples, and β is a suitable constant greater than 1. As the user interacts with the retrieval system, the distance matrix will gradually reflect the distances between the images in semantic space which is consistent with human perception.

3.2 Using Manifold Structure for Image Representation

In the above subsection, we have obtained a distance matrix in semantic space. In this subsection, we discuss how to find the semantic representation for each image in database, while the distances are preserved. Recently, there has been some renewed interest [1][15][11] in the problem of developing low dimensional representations when data arises from sampling a probability distribution on a manifold. To choose a proper mapping algorithm, the following two requirements should be satisfied:

1) Since the image distribution in feature space is highly irregular and inconsistent with human perception, the mapping algorithm must have the locality preserving property.

2) The mapping algorithm should explicitly take into account the manifold structure.

Based on these two considerations, we use Laplacian Eigenmaps [1] to find such a mapping. We first compute the similarity matrix as follows:

$$W_{ij} = \begin{cases} \exp(-B_{ij}/t) & if\ B_{ij} < \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

where t and ε is a suitable constant, and B is the distance matrix obtained in the previous subsection. Note that, the weight matrix has locality preserving property, which is the key feature of Laplaican Eigenmaps.

Suppose $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ is a one-dimensional map of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ in the LE semantic space. A reasonable criterion for choosing a "good" map is to minimize the following objective function under appropriate constraints:

$$\min_{\mathbf{y}} \sum_{i,j} (y_i - y_j)^2 W_{ij}$$

The objective function with our choice of weights W_{ij} incurs a heavy penalty if neighboring points $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far apart. Therefore, minimizing it is an attempt to ensure that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close" then y_i and y_j are close as well. To minimize this objective function, it is equivalent to solve the following eigenvector problem:

$$L\mathbf{y} = \lambda D \mathbf{y}$$

where D is a diagonal matrix, whose entry is column sum (also row sum, since W is symmetric) of matrix W, $D_{ii} = \Sigma_j W_{ji}$. L is called Laplacian matrix, $L = D - W$. Let $\mathbf{y}^{(0)}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}$ be the solutions of the above eigenvector problem, ordered according to their eigenvalues, $\lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_n$. It is easy to show that $\lambda_0 = 0$, and $\mathbf{y}^{(0)} = (1, \ldots 1)$. We leave out $\mathbf{y}_0$ and use the next k eigenvectors for embedding in k-dimensional Euclidean space.

$$\mathbf{x}_i \to \mathbf{z}_i = (\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \cdots, \mathbf{y}_i^{(k)})$$

where $\mathbf{y}_i^{(j)}$ is the ith entry of the eigenvector $\mathbf{y}^{(j)}$. $\mathbf{z}_i$ is a k-dimensional map of image $\mathbf{x}_i$ in the LE semantic space.

In summary, our goal is to find a vector representation (map) in semantic space for each image *in database*. Dimensionality reduction itself is not our goal, though we can make the dimensionality of the LE semantic space much lower than the feature space.

3.3 Learning the Optimal Mapping Function

In the above section, every image in database is mapped into the semantic space. Now, the problem is that, for a new image *outside* the image database, it is unclear how to evaluate its map in the LE semantic space, since we don't have a mapping function. Here we present an approach that applies neural network to approximate the optimal mapping function, which intrinsically distinguishes our framework from previous work [6]. The optimal mapping function f^* is given by minimizing the following cost function:

$$f^* = \arg\min_f \sum_{i=1}^m \left\| f(\mathbf{x}_i) - \mathbf{z}_i \right\|^2$$

where m is the number of images in database. Clearly, this is a multivariate nonparametric regression problem, since there is no *a priori* knowledge about the form of the true mapping function which is being estimated.

In this work, we use radial basis function (RBF) networks, and the standard gradient descent is used as a search technique. The mapping function learned by RBF networks can be represented by

$$f_i(\mathbf{x}) = \sum_{j=1}^h \omega_{ij} G_i(\mathbf{x})$$

where h is the number of hidden layer neurons, $\omega_{ij} \in \mathbf{R}$ are the weights. G_i is the radial function defined as follows:

$$G_i(\mathbf{x}) = \exp(-\frac{\left\| \mathbf{x} - \mathbf{c}_i \right\|^2}{\sigma_i^2})$$

where $\mathbf{c}_i$ is the center for G_i, and σ_i is the basis function width. The k-dimensional mapping in semantic space can be represented as follows:
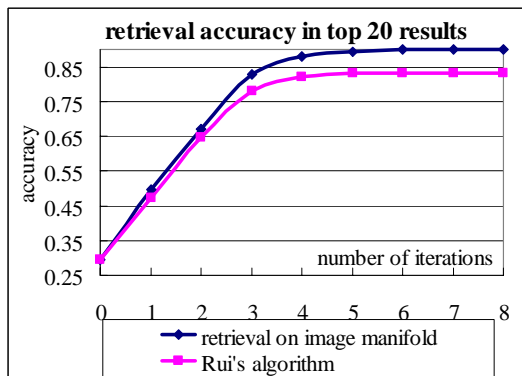
$$\mathbf{x} \to f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_k(\mathbf{x}))$$

where $f = [f_1, f_2, \ldots, f_k]$ is the mapping function. Since the mapping function is approximated by the RBFNN (radial basis function neural network), we call this semantic space RBFNN semantic space.

In summary, the RBF neural network approximates the optimal mapping function from low-level feature space to semantic space. It is trained off-line with the training samples $\{\mathbf{x}_i, \mathbf{z}_i\}$. The computational complexity in retrieval process will be reduced as the dimensionality of the semantic space is reduced. The image representation $f(\mathbf{x}_i)$ in RBFNN semantic space is an approximation of image representation $\mathbf{z}_i$ in LE semantic space, i.e., $f(\mathbf{x}_i) \approx \mathbf{z}_i$. For a new image previously unseen, it can be simply mapped into the RBFNN semantic space by the mapping function f.

4. EXPERIMENTAL RESULTS

In this paper, we focus on image retrieval based on user's relevance feedback to improve the system's short-term and long-term performances. The user can submit a query image either inside or outside the database. The system first computes low-level features of the query image and then maps it into semantic space using the learned mapping function. The system retrieves and ranks the

Figure 2. Comparison of retrieval on image manifold with Rui's algorithm.

Figure 3. Image retrieval performance in low-level feature space, RBFNN semantic space and LE semantic space. The query images are from the image database (*training set*).

images in the database. Then, the user provides his judgment of the relevance of retrieval. With the user's relevance feedback, the system refines the search result iteratively until the user is satisfied. The accumulated relevance feedbacks are used to construct and update the semantic space, as described in Section 3.

We performed several experiments to evaluate the effectiveness of our proposed approaches over a large image dataset. The image dataset we use consists of 3,000 images of 30 semantic categories from the Corel dataset. Each semantic category contains 100 images. The 3,000 images are divided into two subsets. The first subset consists of 2,700 images, and each semantic category contains 90 images. The second subset consists of 300 images, and each semantic category contains 10 images. The first subset is used as training set for learning the optimal mapping function. The second subset is for evaluating the *generalization capability* of our learning framework. A retrieved image is considered correct if it belongs to the same category of the query image. Three types of color features (color histogram, color moment, color coherence) and three types of texture features (tamura coarseness histogram, tamura directionary, pyramid wavelet texture) are used in our system. The combined feature vector is 435-dimensional.

We designed an automatic feedback scheme to model the short-term retrieval process. We only require the user to provide positive examples. At each iteration, the system selects at most 5 correct images as positive examples (positive examples in the previous iterations are excluded from the selection). These automatic generated feedbacks are used as training data to perform short-term learning. To model the long-term learning, we randomly select images from each category as the queries. For each query, a short-term learning process is performed and the feedbacks are used to construct the semantic space. The retrieval accuracy is defined as follows:

$$Accuracy = \frac{relevant\ images\ retrieved\ in\ top\ N\ returns}{N}$$

Four experiments are conducted. The experiment with the new retrieval algorithm on image manifold is discussed in Section 4.1. In Section 4.2, we show the image retrieval performance in the learned semantic spaces. The generalization capability is also evaluated. In Section 4.3 we further test the system's performance in semantic space with different dimensionalities. We compare
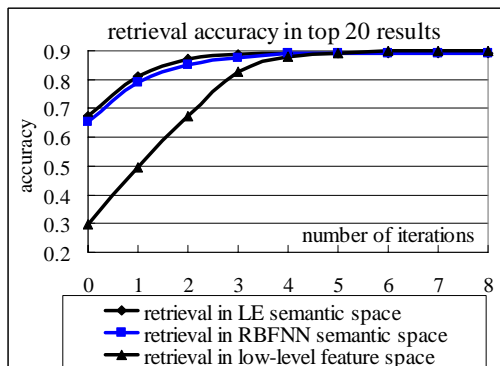
our new algorithm with Rui's algorithm [12] in semantic space in Section 4.4.

4.1 Retrieval on Image Manifold

We compare the performance of our proposed retrieval algorithm on image manifold with the relevance feedback approach described in Rui [12]. We didn't compare it to other image retrieval methods because our primary purpose is to analyze the *geometrical* structure of the image space. Specifically, we aim at comparing the Euclidean structure and manifold structure for data representation in image retrieval. The comparison was made in the low-level feature space with no semantic information involved. Figure 2 shows the experimental result by looking at the top 20 retrievals. As can be seen, our algorithm outperforms Rui's approach. One reason is that the image manifold is possibly highly nonlinear, while Rui's approach can only discover the linear structure.

4.2 Retrieval in Semantic Space

4.2.1 Query Image Inside the Database

As we discussed in Section 3, there are two different semantic spaces, LE semantic space and RBFNN semantic space. One limitation of the LE semantic space is that, it only contains those images in database, i.e., *training set*. It is unclear how to evaluate the map in the LE semantic space for new *test* data. To overcome this limitation, a mapping function f from low-level feature space to high-level semantic space (LE semantic space) is learned by a RBF neural network. That is, the image representation in LE semantic space, z_i, is approximated by $f(x_i)$ which is the image representation in RBFNN semantic space. Intuitively, the retrieval performance in RBFNN semantic space should not be better than that in LE semantic space, since RBFNN semantic space is an approximation of LE semantic space.

Figure 3 shows the retrieval performance in low-level feature space, LE semantic space and RBFNN semantic space. Our new retrieval algorithm on image manifold is used. We use the training set (2700 images) as the image database. We first conduct the experiment in low-level feature space. As the previous experiment in Section 4.1, we randomly choose 20% of images in each semantic class as queries to perform the retrieval. The user's relevance feedbacks are used to learn a LE semantic space, a RBFNN
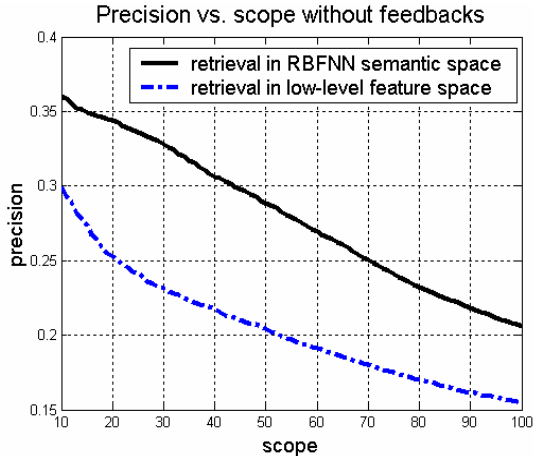
Figure 4. Image retrieval performance in low-level feature space and RBFNN semantic space. The query images are *outside* the database.

Figure 5. The initial retrieval accuracy (no feedback is provided) in RBFNN semantic space with different dimensionalities.
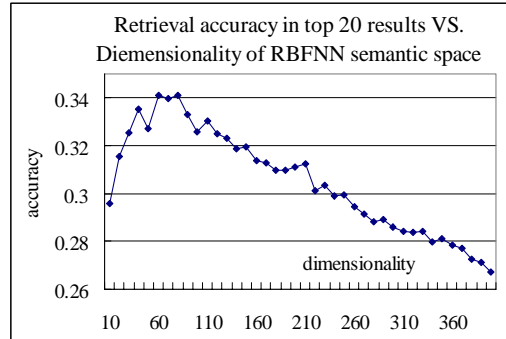
semantic space, as well as a mapping function from 435-dimensional low-level feature space to 30-dimensional high-level semantic space. We will discuss how to determine the intrinsic dimensionality of the semantic space in the next section. As can be seen, the retrieval performance in semantic spaces is much better than that in low-level feature space. The performance difference is especially significant at the 0th retrieval when no user's relevance feedback is provided, which is our primary goal. That is, the semantic representation of the query image can be learned by the RBF neural network f which is trained by previous users' interactions with the system. In fact, in the real world, if the initial retrieval result is too bad, the user might lose his interest to provide feedbacks.

Another observation is that, the retrieval performances in LE semantic space and RBFNN semantic space are almost the same. This means that the optimal mapping function f^* can be accurately approximated by the RBF neural network f.

4.2.2 *Query Image outside the Database --- Generalization Capability Evaluation*

While using RBF neural network to solve the regression problem, a key issue is its *generalization* capability. Generalization refers to the neural network producing reasonable outputs for inputs not encountered during training. To evaluate the generalization capability of our model, the 300 images (*testing set*) are used as queries outside the image database (*training set*) for testing. These images have no semantic representations in LE semantic space, but we can obtain their semantic representations in RBFNN semantic space by the mapping function f. Since our intention is to evaluate the generalization capability of our model, the initial retrieval result is especially important when no feedbacks are provided. The precision-scope curves are shown in Figure 4. As can be seen, the retrieval in RBFNN semantic space outperforms that in low-level feature space. This means that the semantic representation of the previously unseen images can be accurately learned by the RBF neural network.

4.3 Retrieval in Semantic Space with Different Dimensionalities

One issue of learning a semantic space is how to estimate its intrinsic dimensionality. Even though the dimensionality of low-level feature space is normally very high, the dimensionality of semantic space is much lower.

In this section, we evaluate the retrieval performance in semantic spaces with different dimensionality. As before, the 300 images outside the image databases are used as the query images. Both the images in database and the query images are mapped into RBFNN semantic space by the mapping function. Figure 5 shows the results. As can be seen, the optimal dimensionality is closely related to the number of semantic classes in the database. This observation coincides with that obtained in [6]. If the image database administrator has a prior knowledge about this number, it can be used as a guideline to control the dimensionality of the semantic space. The system reaches the best performance (in terms of accuracy and efficiency) when the dimensionality of the semantic space is close to the number of semantic classes. Further compression of the semantic space will start to cause information loss and decrease the retrieval accuracy.

4.4 Comparing Different Retrieval Algorithms in Semantic Space

In previous two subsections, we have evaluated the retrieval performance in semantic space using our retrieval algorithm. It is interesting to see how Rui's algorithm [12] performs in semantic space. Figure 6 shows the retrieval results using our retrieval algorithm and Rui's algorithm. We use the same image database and the same query images as in Section 4.2.1. The retrieval is conducted in RBFNN semantic space rather than the feature space. As can be seen, Rui's algorithm works almost the same as our algorithm. It is important to note that the baseline performance in semantic space is much higher than that in low-level feature space. This observation confirms our previous intuition that the semantic space gets more and more "regular" (flat and linear) as the user's relevance feedback is incorporated. To be specific, in the semantic space, the geodesic distances are almost equal to the Euclidean distances (see Figure 1(d)). Hence the Riemannian structure of the image manifold can be inferred from the Euclidean structure of the semantic space.

5. CONCLUSIONS

In this paper, under the assumption that the data lie on a submanifold hidden in a high dimensional feature space, we developed an
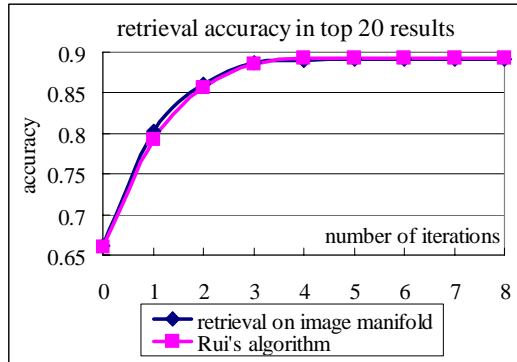
Figure 6. The comparison of our retrieval algorithm with Rui's algorithm in semantic space. The performances of these two algorithms are very close. It shows that the semantic space is Euclidean (flat).

algorithmic framework to learn the mapping between low-level image features and high-level semantics. It utilizes relevance feedback to enhance the performance of image retrieval system from both short- and long-term perspectives. This framework gives a solution to the two fundamental problems in image retrieval: *how to judge similarity and how to represent an image*.

To solve the first problem, the proposed retrieval algorithm on image manifold uses the geodesic distance rather than Euclidean distance as the similarity measure between images. It takes into account the Riemannian structure of the image manifold on which the data may possibly reside.

To solve the second problem, two semantic spaces, LE semantic space and RBFNN semantic space, are learned from user's relevance feedback. A mapping function is approximated by a RBF neural network. The semantic space gives a Euclidean representation of the Riemannian image manifold.

Several questions remain unclear:

1. We do not know how often and in which particular empirical contexts, the manifold properties are crucial to account for the underlying topology of image data. While the results in this paper provide some indirect evidence for this, there still seems to be no convincing proof that such manifold structures are actually present.

2. Secondly, and most intriguingly, while the notion of semantic space is a very appealing one, the properties of the *true* mapping from low-level feature space to high-level semantic space remains unclear. It is unclear whether the true mapping is one-to-one, or many-to-one, since intuitively two different images might have totally the same semantics. The mapping function is learned in a statistical sense. Though the experiments show its strong generalization capability, it still remains unclear in theory.

REFERENCES

[1] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation", *Advances in Neural Information Processing Systesms*, 2001.

[2] B. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum, "Graph approximations to geodesics on embedded manifolds", Technical report, Stanford University, December 2000

[3] E. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machine". *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, No. 1, Jan. 2003.

[4] X. He and P. Niyogi, "Locality Preserving Projections", in *Advances in Neural Information Processing Systems 16*, Vancouver, Canada, 2003.

[5] X. He, D. Cai, H. Liu and W.-Y. Ma, "Locality Preserving Indexing for Document Representation", in *ACM SIGIR conference on Information Retrieval*, Sheffield, 2004.

[6] X. He, O. King, W.-Y. Ma, M.-J. Li, and H.-J. Zhang, "Learning a semantic space from user's relevance feedback for image retrieval", *IEEE Trans. on Circuit and System for Video Technology,* Jan, 2003.

[7] X. He, S. Yan, Y. Hu and H.-J. Zhang, "Learning a locality preserving subspace for visual recognition", in *Proc. IEEE Conf. on Computer Vision,* Nice, France, 2003.

[8] Y. Ishikawa, R. Subramanya and C. Faloutsos, "MindReader: query databases through multiple examples", *24th Conf. on Very Large Databases,* New York, 1998.

[9] W.-Y. Ma and B. S. Manjunath, "Netra: A toolbox for navigating large image databases", *Multimedia System Journal*, vol. 7, pp. 184-198, 1999.

[10] W. Matusik, H. Pfister. M. Brand, and L. McMillan, "A data-driven reflectance model", in *Proc. of SIGGRAPH*, 2003.

[11] S.T. Roweis, and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", *Science*, vol 290, 22 December 2000.

[12] Y. Rui and T. S. Huang, "Optimizing learning in image retrieval", in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Hilton Head, SC, June 2000.

[13] H.S. Seung and D. Lee, "The manifold ways of perception", *Science*, vol 290, 22 December 2000.

[14] J. Smith and S.F. Chang, "VisualSEEk: A fully automatic conent-based image query system", *ACM Multimedia*, 1996.

[15] J.B. Tenenbaum, V.D. Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction", *Science*, Vol 290, 22 December 2000.

[16] K. Tieu and P. Viola, "Boosting image retrieval", in *Proc. IEEE Conf. on Computer Vision and Pattern Recognitino*, Hilton head, SC, June 2000.

[17] S. Tong and E. Chang, "Support vector machine active learning for image retrieval", in *Proc. ACM Multimedia 2001*, Ottawa, Canada, 2001.

[18] N. Vasconcelos and A. Lippman, "Learning from user feedback in image retrieval systems", *Advances in Neural Information Processing Systems*, Denver, Colorado, 1999.

[19] J. Wang and J. Li, "Learning-based linguistic indexing of pictures with 2-D MHMMs", in *Proc. ACM Multimedia*, pp. 436-445, Juan Les Pins, France, 2002.

[20] X. S. Zhou and T.S. Huang, "Comparing Discriminating Transformations and SVM for Learning during Multimedia Retrieval", in *Proc. ACM Multimedia* 2001, Ottawa, 2001.

[21] L. Zhu, A. Rao and A. Zhang, "A theory of keyblock-based image retrieval", *ACM Trans. on Information Systems*, vol. 20, No. 2, pp. 224-257, 2002.