Full Paper in ACM Multimedia, Santa Barbara, CA, October 23-27, 2006.

Learning Image Manifolds by Semantic Subspace Projection

Jie Yu
Department of Computer Science
University of Texas at San Antonio
San Antonio, TX 78249, USA
(+01) 210-458-7449
jyu@cs.utsa.edu

Qi Tian
Department of Computer Science
University of Texas at San Antonio
San Antonio, TX 78249, USA
(+01) 210-458-5165
qitian@cs.utsa.edu

ABSTRACT

In many image retrieval applications, the mapping between high-level semantic concept and low-level features is obtained through a learning process. Traditional approaches often assume that images with same semantic label share strong visual similarities and should be clustered together to facilitate modeling and classification. Our research indicates this assumption is inappropriate in many cases. Instead we model the images as lying on non-linear image subspaces embedded in the high-dimensional space and find that multiple subspaces may correspond to one semantic concept. By intelligently utilizing the similarity and dissimilarity information in semantic and geometric (image) domains, we find an optimal Semantic Subspace Projection (SSP) that captures the most important properties of the subspaces with respect to classification. Theoretical analysis proves that the well-known Linear Discriminant Analysis (LDA) could be formulated as a special case of our proposed method. To capture the semantic concept dynamically, SSP can integrate relevance feedback efficiently through incremental learning. Extensive experiments have been designed and conducted to compare our proposed method to the state-of-the-art techniques such as LDA, Locality Preservation Projection (LPP), Local Linear Embedding (LLE), Local Discrimiant Embedding (LDE) and their semi-supervised algorithms. The results show the superior performance of SSP.

Categories and Subject Descriptors

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

General Terms

Algorithms, Theory, Performance, Experimentation, Measurement.

Keywords

Semantic Subspace Projection, Image Retrieval, Relevance Feedback, Subspace Learning, Principal Component Analysis, Linear Discriminat Analysis

1. INTRODUCTION

With the development of digital imaging technology, more and more information nowadays is conveyed in the form of digital images or video clips. The rich context of an image makes the understanding of its semantic meaning very difficult. Content-based image retrieval (CBIR) aims at automatically retrieving the images of user interest from large database based on their visual content [1]. The user interest could be summarized by a high-level semantic concept while the visual content of the images are usually represented by low-level features such as color, texture and shape. Machine learning techniques are used to bridge the gap between the semantic concept and image features. In that process the user gives semantic information on a few sample images by labeling them. Statistical model of the images with same semantic label could be estimated based on the labeled training set.

Because the dimensionality of feature space is usually very high, direct model estimation in the high dimensional feature space is easy to fail. Dimension reduction is used to map the original space to a low dimensional space. The concept learning is conducted in that projected space. In that process if insufficient samples are used for training purpose a projection that facilitates the learning is difficult to obtain. Thus the small sample set problem along with the high dimensionality problem become two major challenges for image retrieval.

Small Sample Set: In CBIR, a set of samples with categorical information are used to train a classifier. Because labeling the training samples requires human interference and could be computationally expensive, the size of the training set is often very small. In that case the learning process tends to bias to the training set and overfitting could occur.

High Dimensionality: In many data analysis application, the observed data have very high dimensionality. Specifically the images in CBIR are represented by image feature vector whose dimensionality ranges from tens to hundreds in most cases. Traditional statistical approaches have difficulties in modeling data directly in such a high dimensional space.

Some techniques have been proposed to alleviate the two problems. For the small sample set problem, researchers have been using unlabeled data along with the labeled data to avoid overfitting. However our research in [2] shows that the unlabeled data and labeled data must be from the same statistical source to improve the performance. Otherwise using unlabelled data may deteriorate the classification. For the high dimensionality problem, it is almost a common practice to conduct dimension reduction to find a compact representation of data in a low dimensional space. Note that the small sample set problem are related to the high dimensionality problem in that, when the data dimensionality is high, sample set size usually needs to be relatively large to get a accurate model of the data. In that sense the dimension reduction techniques can also alleviate the small sample set problem indirectly since data modeling is usually applied on the projected data with low dimensionality. Traditional techniques, such as Principal Component Analysis (PCA) [3] and Linear Discriminat Analysis (LDA) [4,5], are widely applied in many applications. For a classification task such as CBIR, LDA is often preferred because it incorporates class information and discovers the most discriminant features in the projected space. However, research has indicated that both PCA and LDA assume image data are from certain distribution model (in most cases Gaussian or Gaussian Mixture). In many applications, this assumption is inappropriate.

Recently more and more attention has been drawn on modeling the data as lying on a subspace which is embedded in the high dimensional space [6-11,16]. The intrinsic structure of the subspace could be discovered and preserved in a low dimensional space by using subspace learning techniques. Because the global structure of the subspace is inferred from the local neighborhood information, the manifold learning techniques do not assume all the data are from a specific distribution as in PCA and LDA.

The rest of the paper is organized as follows. In Section 2 we analyze the relation between semantic class and geometric subspaces in image retrieval and discuses on how to use similarity and dissimilarity properly in learning the semantic concept. Section 3 introduces our proposed method Semantic Subspace Projection (SSP) that intelligently uses semantic information on subspace learning to capture the most important properties for classification. Section 4 shows the experiment results on benchmark data set and image retrieval applications. Conclusions and future work are discussed in Section 5.

2. SEMANTIC CLASS AND GEOMETRIC SUBSPACES IN IMAGE RETRIEVAL

2.1 Semantic Class and Geometric Subspaces

As we mentioned in the introduction, in image retrieval the images are often represented by feature vectors, which correspond to visual content such as color, texture and shape. The low-level image features construct a high dimensional geometric space while the semantic label of an image is associated with the high-level human concept. Our method tries to bridge the gap between the semantic domain and geometric space by adding semantic correlation of data.

In our approach, the images are modeled as data lying on subspaces in the geometric space and its structure can be captured by preserving the geometric neighborhood (local patch) information. Compared to PCA and LDA, the subspace model is more robust because the global structure of the data does not have to be Gaussian. The semantic concept can be learned by capturing the structure of the geometric subspaces.

Traditional techniques such as PCA and LDA assume that all the images in a semantic class share strong visual similarity and can be modeled as lying on a single geometric subspace. Figure 1 (a) illustrates the one-to-one relation between semantic class and geometric subspace. The semantic concepts corresponding to such semantic classes are often simple, such as “banana” and “sunset”.

However, the real scenarios in most CBIR applications are very complicated, which make the one-to-one assumption of semantic class and geometric subspace invalid. The reason lies in two folds:

1) In most CBIR systems, the users label the images from various sources as relevant to their interest (positive) or irrelevant to their interest (negative). Naturally the subjects of images from negative class are usually different and hard to define. Thus multiple geometric subspaces often co-exist within the negative semantic classes.

2) More importantly, the human judgment on the semantic subject of an image may be from non-visual knowledge. The user may group images with dramatically different visual contents into the same semantic class. For instance one who is interested in images of fruit may result in labeling strawberry and banana as positive in spite of the fact that they are visually quite dissimilar. In that case the images within same (positive or negative) semantic class may be better modeled by multiple geometric subspaces.

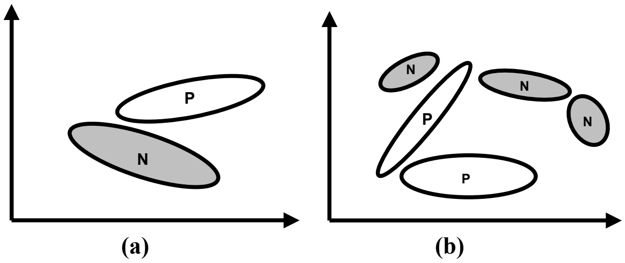


Fig. 1: Illustration of semantic class and geometric subspaces (P: Positive, N: Negative)

Figure 1(b) gives us an example of that one-to-multiple relation between semantic class and geometric subspaces could exist among negative and/or positive samples. The fact that one semantic class may contain multiple geometric subspaces constructs a more challenging learning problem. Supervised and unsupervised approaches [12,13,28] have been proposed to alleviate the problem. However they often assume the data distribution is Gaussian which is inappropriate in some cases. Besides, in those approaches the number of subspaces has to be precisely estimated which could be a difficult and computationally expensive. With the subspace model, our proposed method does not have to model the global distribution. Instead it aims at learning the global structure of the semantic manifold from locality point of view and inexplicitly generalizes the semantic concept.

2.2 Similarity and Dissimilarity

The visual resemblance between images can be defined as similarity and dissimilarity in the geometric space. Research on similarity and dissimilarity measure for image retrieval has been found in literature [14]. However the relation between geometric similarity and semantic similarity is unclear. We found that for an image query the user’s judgment on the semantic similarity and dissimilarity between images may not be solely based the geometric information. A classifier will predict the semantic relation between two images based on their geometric correlation. It is obvious that we can deduct all four possibilities for the combination of geometric and semantic similarity relation:

Let the query image be *q*, while one image from the database be *img*. The semantic relation between *q* and *img* is denoted as

$R^S(q, img) = \{Similar, Dissimilar\}$ while the geometric/visual relation predicted by a classifier is denoted as $R^G(q, img) = \{Similar, Dissimilar\}$. Then we have

(i) $R^G(q, img) = Similar \Rightarrow R^S(q, img) = Similar$

(ii) $R^G(q, img) = Dissimilar \Rightarrow R^S(q, img) = Dissimilar$

(iii) $R^G(q, img) = Similar \Rightarrow R^S(q, img) = Dissimilar$

(iv) $R^G(q, img) = Dissimilar \Rightarrow R^S(q, img) = Similar$

The case 1 and 2 can be easily accepted and interpreted. One may find two images visually/geometric similar and fall in the same category, which is described by case 1. On the other hand it is easy to justify case 2 when the user find two images as semantic dissimilar then there must be some visual difference that should be learnt by a classifier. Case 3 indicates that although one classifier find two images visually similar, they are semantically irrelevant. It is the case that a well-trained classifier should avoid. Case 4 is where traditional approaches will fail. Because of the multiple-to-one relation between geometric subspaces and semantic class, it is possible that two images are on different geometric subspaces within the same semantic manifold. In this sense they are geometrically dissimilar while semantically related. In our approach we carefully utilize the semantic information from user and enable our algorithm could handle complex query task illustrated in case 4.

Based on the above analysis, we conclude two guidelines for intelligent use of the semantic information user provides:

(i) It is safe to claim two images are visually dissimilar if the user label them as semantically irrelevant to each other because the user must make the judgment based on some visual difference between the images.

(ii) If the user assigns two images to same semantic category, it does not suggest that they must be visually similar because that decision may be based on some non-visual prior knowledge.

From the above observations we can find the semantic dissimilarity information is also very important in learning complex concept while the semantic similarity may not correspond to the visual similarity. It is desirable to use both semantic similarity and dissimilarity information for self-discovery of the geometric subspaces.

2.3 Related Work

To facilitate data exploration, researchers have been trying to capture the structure of correlated data group by mapping the original high-dimensional space to a low-dimensional one. PCA tries to find an optimal projection that keeps the most descriptive features. Because PCA is an unsupervised approach and can't take semantic information. It is often substituted by LDA in classification applications. LDA aims at making the images with same semantic label cluster and the images from different semantic categories separate. However LDA is easy to suffer from the small sample set problem since it assumes Gaussian distribution and uses sample-based estimation.

In recent years several subspace learning techniques have been proposed to apply in this field. Instead of assuming data are from a particular distribution, they could be modeled as lying on a non-linear low-dimensional subspace embedded in the high-dimensional space. The global structure of such a subspace could be inferred by gathering local information of every neighborhood on that subspace. Local Linear Embedding (LLE) [6] assumes the local patch can be approximated by a hyperplane and one sample could be approximated by linear combination of its neighbors. This linear relation between neighbors is invariant to scaling and affine transformation. In LLE the global structure is considered as well captured in the projected space when the neighborhood correlation in the original space is best preserved in the low dimensional space. ISOMAP [7] starts from the observation that global structure of a manifold is non-Euclidean and only local patches of the manifold can be safely modeled as Euclidean. Consequently it proposes to use geodesic distance to substitute Euclidean distance to reflect the non-linear structure of the subspace. Locality Preserving Projection (LPP) [9] differs from LLE in that it treats the neighborhoods as clusters. Instead of solving for the linear reconstruction correlations between neighbors, LPP assumes the locality information is preserved as long as the neighbors in the original space keep their neighborhood relationship in the projected space. The optimal projection found by LPP is the one that makes neighbor data close to each other in the projected space. It is clear that these three techniques are all unsupervised. Although they have been successfully applied in many applications, they face some difficulties in classification task in that semantic information could be better integrated into learning the user interest.

Supervised and semi-supervised approaches based on the manifold learning techniques are further proposed since they are more desirable with respect to classification. Supervised LLE (S-LLE) [15] incorporates user provided information by tuning a parameter α to control the influence of semantic labeling on the geometric structure it learns. Incremental Semi-supervised LPP (I-LPP) [10] tries to use user feedback as semantic relation to override the geometric neighborhood relation graph. Theoretical analysis proves that it can converge to minimizing the well-known within-class scatter matrix (S_w). S-LLE and I-LPP do not use dissimilarity information to make samples from different semantic classes apart in the projected space. Local Discrimiant Embedding (LDE) [16] and its semi-supervised version Augmented Relation Embedding (ARE) [17] resembles our work in that it utilizes geometric and semantic similarity information in a local neighborhood by clustering only samples with same semantic labels and from the same neighborhood. However we find the semantic dissimilarity should be better used in learning a semantic manifold. In LDE only samples from different classes but within the same geometric neighborhood will be projected far away from each other. However problems could arise due to the fact that two samples from different classes that are well apart in the original space could be projected as close to each other in the low-dimensional space. If no penalty is given to properly handle this case, the between-class separation could be deteriorated. Furthermore similarly as I-LPP the semi-supervised LDE (ARE) tries to override geometric similarity with semantic similarity when re-training the classifier based on user feedback. As we indicate in Section 2.1, the semantic similarity may not arise from geometric similarity and can not be used directly in learning the semantic concept. Based on the above discussion, we are motivated to propose a novel method that could learn semantic manifold intelligently.

3. SEMANTIC SUBSPACE PROJECTION

3.1 Methodology

Inspired by the success of the state-of-the-art techniques, we propose a novel framework that could discover the structure of semantic manifold in respect to classification and overcome the weakness of the state-of-the-art techniques. As we indicate in Section 2.1, there is a gap between geometric structure and semantic similarity. To better bridge them, we propose a novel supervised method called *Semantic Subspace Projection* (SSP). It could make use of user provided semantic information and capture the structure of each geometric subspace. A brief introduction of the technique is as follows:

Because the semantic dissimilarity between images must arise from some visual differences, those images from different semantic classes should be as separated as possible in the new space. The pair-wise Semantic Dissimilarity information between two samples $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ can be stored in a matrix ***SD*** as in equation (1). It uses supervised information to refine the geometric structure graph.

$$SD_{ij} = \begin{cases} 1 \text{ if } \boldsymbol{x}_{\text{i}}, \boldsymbol{x}_{\text{j}} \notin \text{same semantic class} \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

Instead of assuming that each semantic class contains only one subspace, we try to discover the global structure of the data by preserving locality information in a projected space. The local geometric information is first captured by constructing a *K* Nearest Neighbor (*KNN*) graph:

$$GeoSim_{ij} = \begin{cases} 1 \text{ if } \boldsymbol{x}_{\text{i}}, \boldsymbol{x}_{\text{j}} \text{ both within } K \text{ NN} \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

Considering that some visually similar, that is, geometrically close, image data may be from semantically different classes, we incorporate the semantic information into the geometric structure discovered and construct a new Geometric-Semantic Similarity Matrix:

$$GSSim_{ij} = \overline{SD}_{ij} \cdot GeoSim_{ij} \tag{3}$$

where " $\cdot$ " denotes the element-wise Boolean product operation.

Because each neighborhood may contain different number of samples, the Geometric-Semantic Similarity Matrix is normalized for each row.

$$GSSim_{ij} = GSSim_{ij} / \sum_j GSSim_{ij} \tag{4}$$

In the above matrix, the semantic dissimilarity is considered for classification purpose and geometric neighborhood information, along with semantic similarity information, is used to discover image subspaces.

Let *W* be a projection that maps any sample $\boldsymbol{x_i}$ in original space to a corresponding sample $\boldsymbol{y_i}$ in a lower dimension space.

$$\boldsymbol{y}_i = W^T \boldsymbol{x}_i \tag{5}$$

Obviously in the local neighborhood of a sample $\boldsymbol{x_i}$, the mean can be estimated as:

$$m_i = \sum_j \boldsymbol{x}_j GSSim_{ij} \tag{6}$$

After projection, the mean can be calculated from (5) and (6):

$$m_i^{(y)} = \sum_j \boldsymbol{y}_j GSSim_{ij} \tag{7}$$

Intuitively we propose a new cost function for finding the optimal projection:

$$\begin{aligned} W &= \arg\max_W \frac{|\sum_{i,j}(m_i^{(y)} - m_j^{(y)})(m_i^{(y)} - m_j^{(y)})^T SD_{ij}|}{|\sum_{i,j}(y_i - y_j)(y_i - y_j)^T GSSim_{ij}|} \\ &= \arg\max_W \frac{|\sum_{i,j} W^T(m_i - m_j)(m_i - m_j)^T SD_{ij} W|}{|\sum_{i,j} W^T(x_i - x_j)(x_i - x_j)^T GSSim_{ij} W|} \end{aligned} \tag{8}$$

In (8) the numerator corresponds to the separation between local neighborhood means from different semantic classes. The different semantic classes are more far away from each other when the numerator gets larger. Note that we do not use inter-class distances for numerator as in LDA because there may be multiple subspaces in one semantic class. The denominator aims at preserving global structure by clustering the samples within the same local patches. It models the data better because subspaces structure is preserved. Consequently the classification in the projected space will be enhanced, especially when there are multiple subspaces within one semantic class. Note that by using *GSSim*, only samples that are both semantic and geometric similar are clustered together unlike in existing approaches where all samples from same semantic classes are forced together. That may result in better discovery of the global class structure because all geometric subspacees' structure is preserved. By maximizing the ratio of the numerator and denominator we can find an optimal projection that makes the subspaces from different semantic classes more separate from each other and the samples within same geometric and semantic neighborhood clustered together.

We denote semantic Dissimilarity Scatter Matrix as follows:

$$S_{Diss} = \sum_{i,j}(m_i - m_j)(m_i - m_j)^T SD_{ij} \tag{9}$$

The Geometric-Semantic Similarity Scatter Matrix can be defined as:

$$S_{GS-Sim} = \sum_{i,j}(x_i - x_j)(x_i - x_j)^T GSSim_{ij} \tag{10}$$

Consequently the optimal projection can be obtained by solving the following optimization problem:

$$\begin{aligned} & W = \arg\max_W \frac{|W^T S_{Diss} W|}{|W^T S_{GS-Sim} W|} \\ & \Rightarrow W = \underset{\max}{eig}(S_{Diss} S_{GS-Sim}^{-1}) \end{aligned} \tag{11}$$

The optimal projection *W* consists of the eigenvectors corresponding to the largest eigenvalues of $S_{Diss} S_{GS-Sim}^{-1}$. It is worth mentioning that SSP is general and, consequently, any classification method could benefit from our proposed method by applying on the projected data instead of on the originally high dimensional data.

Compared to other subspace learning techniques, SSP is novel in that:

(i) We consider not only preserving the subspaces' structure after the projection but also separating all samples from different classes, which is omitted in LLE, LPP and their semi-supervised algorithms. For a classification task the latter obviously can not be neglected.

(ii) SSP uses semantic similarity and geometric similarity as joint constraints to define local neighborhood while in I-LPP and ARE semantic similarity overrides geometric information. In the case of multiple subspaces corresponding to one semantic class as we discussed in Section 2.1, visually dissimilar images may be forced to project close to each other, which is unnecessary and easy to fail. However this problem does not exist for our method. The reason is that SSP tries to capture the structure of subspaces by clustering only samples within a geometric and semantic local neighborhood and consequently no effort will be put to cluster samples from different subspaces. Further discussion on this will be given later in Section 3.3.

(iii) In SSP, all the samples from different classes are projected apart from a global point of view while in LDE only samples from different class but within the same neighborhood will be separated from each other. Problems could arise in their approach due to the fact that two samples from different classes, which are well apart in the original space, could be projected as close to each other in the low-dimensional space. If no penalty is given to properly handle that case, the inter-class separation could be deteriorated.

3.2 Relation to LDA

Linear discriminant analysis has been widely used in image retrieval applications. It assumes that one subspace from Gaussian distribution corresponds to one semantic concept. Our proof below shows that SSP reduces to LDA when the neighborhood is defined as large enough to cover all images from the same semantic class.

If we have *N* training samples, the above condition can be satisfied when $K \geq N$. In that case the Semantic Dissimilarity matrix *SD* is same as in (1) while no geometric neighborhood information can be learned because *GeoSim* is an all 1's matrix as shown in (12).

$$GeoSim_{ij} = 1 \text{ for any pair of images} \tag{12}$$

Since $GSSim = \overline{SD} \cdot GeoSim$, we have

$$GSSim_{ij} = \begin{cases} \dfrac{1}{n_l} if\ \boldsymbol{x}_i, \boldsymbol{x}_j \in semantic\ class\ l \\ 0\ otherwise \end{cases} \tag{13}$$

where we denote the index of semantic class as *l* and the number of samples in class *l* as n_l.

Instead of using semantic information along with geometric information for subspace learning, here the semantic information overrides the geometric information. Consequently for each sample $\boldsymbol{x}_i$, its subspace mean m_i becomes its class mean $m^c_{L(i)}$.

$$m_i = \sum_j \boldsymbol{x}_j GSSim_{ij} = \sum_{L(j)=L(i)} \boldsymbol{x}_j / n_{L(i)} = m^c_{L(i)} \tag{14}$$

where we denote the label of sample $\boldsymbol{x}_i$ as $L(i)$.

Thus the Dissimilarity Scatter Matrix converges to Between-Class Scatter Matrix. Proof is shown below.

Starting with (9) and using (1), we can find $SD_{ij} = 1$ *iff L(i) ≠ L(j)* . Thus we have

$$\begin{aligned} S_{Diss} &= \sum_{i,j,L(i)\neq L(j)_j} (m^c_{L(i)} - m^c_{L(j)})(m^c_{L(i)} - m^c_{L(j)})^T \\ &= \sum_{i,j,L(i)\neq L(j)} (m^c_{L(i)} m^c_{L(i)}{}^T - m^c_{L(j)} m^c_{L(i)}{}^T - m^c_{L(i)} m^c_{L(j)}{}^T + m^c_{L(j)} m^c_{L(j)}{}^T) \end{aligned} \tag{15}$$

In equation (15), the sum of class means over each sample could be rewritten in the form of the sum of class means over the number of samples of each class as below.

$$\sum_{i,j,L(i)\neq L(j)} m^c_{L(i)} m^c_{L(i)}{}^T = \sum_i \sum_{l,L(i)\neq l} n_l m^c_{L(i)} m^c_{L(i)}{}^T \tag{16}$$

Similarly we have:

$$\sum_{i,j,L(i)\neq L(j)} m^c_{L(j)} m^c_{L(i)}{}^T = \sum_i \sum_{l,L(i)\neq l} n_l m^c_l m^c_{L(i)}{}^T \tag{17}$$

$$\sum_{i,j,L(i)\neq L(j)} m^c_{L(j)} m^c_{L(j)}{}^T = \sum_i \sum_{l,L(i)\neq l} n_l m^c_l m^{cT}_l \tag{18}$$

Thus equation (16) could be written as:

$$S_{Diss} = \sum_i \begin{array}{l} (\sum_{l,l\neq L(i)} n_l m^c_{L(i)} m^c_{L(i)}{}^T - \sum_{l,l\neq L(i)} n_l m^c_l m^c_{L(i)}{}^T - m^c_{L(i)} \sum_{l,l\neq L(i)} n_l m^{cT}_l \\ + \sum_{l,l\neq L(i)} n_l m^c_l m^{cT}_l) \end{array} \tag{19}$$

If we denote the grand mean of all samples as:

$$m_G = \sum_i x_i / N = \sum_l n_l m^c_l / N \tag{20}$$

We have the following equation:

$$\sum_{l,l\neq L(i)} n_l m^c_l = N m_G - n_{L(i)} m^c_{L(i)} \tag{21}$$

Thus equation (19) can be further rewritten in the following way

$$\begin{aligned} S_{Diss} &= \sum_i (\sum_{l\neq L(i)} n_l m^c_{L(i)} m^c_{L(i)}{}^T - (N m_G - n_{L(i)} m^c_{L(i)}) m^c_{L(i)}{}^T - m^c_{L(i)} (N m_G{}^T - n_{L(i)} m^c_{L(i)}{}^T) \\ &\quad + \sum_{l\neq L(i)} n_l m^c_l m^{cT}_l) \\ &= \sum_i (N m^c_{L(i)} m^c_{L(i)}{}^T - N m_G m^c_{L(i)}{}^T - N m^c_{L(i)} m_G{}^T + \sum_l n_l m^c_l m^{cT}_l) \\ &= 2N(\sum_l n_l m^c_l m^{cT}_l - N m_G m_G{}^T) \end{aligned} \tag{22}$$

In traditional LDA, the between-class scatter matrix can be rewritten as follows:

$$\begin{aligned} S_B &= \sum_l n_l (m^c_l - m_G)(m^c_l - m_G)^T \\ &= \sum_l (n_l m^c_l m^{cT}_l - n_l m^c_l m_G{}^T - n_l m_G m^{cT}_l + n_l m_G m_G{}^T) \end{aligned}$$

$$= \sum_l n_l m_l^c m_l^{cT} - N m_G m_G{}^T - N m_G m_G{}^T + N m_G m_G{}^T$$

$$= \sum_l n_l m_l^c m_l^{cT} - N m_G m_G{}^T$$

$$= \frac{1}{2N} S_{Diss} \tag{23}$$

Similarly we can prove $S_{GS-Sim} = \frac{1}{2} S_W$ below, when the geometric neighborhood is overridden by semantic class labels.

$$\begin{aligned}
S_{GS-Sim} &= \sum_{i,j} (x_i - x_j)(x_i - x_j)^T GSSim_{ij} \\
&= \sum_i \sum_{L(j)=L(i)} (x_i - x_j)(x_i - x_j)^T / n_{L(i)} \\
&= \sum_i (\sum_{L(j)=L(i)} (x_i x_i{}^T - x_j x_i{}^T - x_i x_j{}^T + x_j x_j{}^T) / n_{L(i)}) \\
&= \sum_i (n_{L(i)} x_i x_i{}^T - n_{L(i)} m_{L(i)}^c x_i{}^T - n_{L(i)} x_i m_{L(i)}^c{}^T + \sum_{L(j)=L(i)} x_j x_j{}^T) / n_{L(i)} \qquad (24) \\
&= \sum_l (\sum_{L(i)=l} (n_l x_i x_i{}^T - n_l m_l^c x_i{}^T - n_l x_i m_l^{cT} + \sum_{L(j)=L(i)=l} x_j x_j{}^T) / n_l) \\
&= \sum_l ((n_l \sum_{L(i)=l} x_i x_i{}^T - n_l m_l^c n_l m_l^{cT} - n_l n_l m_l^c m_l^{cT} + n_l \sum_{L(j)=l} x_j x_j{}^T) / n_l) \\
&= 2 \sum_l (\sum_{L(i)=l} x_i x_i{}^T - n_l m_l^c m_l^{cT})
\end{aligned}$$

The within-class scatter matrix is proved to be proportional to S_{Gs-Sim} as below:

$$\begin{aligned}
S_W &= \sum_i (x_i - m_{L(i)}^c)(x_i - m_{L(i)}^c)^T \\
&= \sum_l \sum_{L(i)=l} (x_i x_i{}^T - m_l^c x_i{}^T - x_i m_l^{cT} + m_l^c m_l^{cT}) \\
&= \sum_l (\sum_{L(i)=l} x_i x_i{}^T - n_l m_l^c m_l^{cT} - n_l m_l^c m_l^{cT} + n_l m_l^c m_l^{cT}) \qquad (25) \\
&= \sum_l (\sum_{L(i)=l} x_i x_i{}^T - n_l m_l^c m_l^{cT}) \\
&= \frac{1}{2} S_{GS-Sim}
\end{aligned}$$

Thus we have

$$W_{SSP} = \arg\max_W \frac{|W^T S_{Diss} W|}{|W^T S_{GS-Sim} W|} = \arg\max_W \frac{|W^T S_B W|}{|W^T S_W W|} = W_{LDA} \tag{26}$$

when the neighborhood size is large enough (*K*>*N*-*1*).

From (26) we can conclude that, LDA becomes a special case of SSP where each class is assumed to contain only one subspace. According to our discussion in Section 2, that assumption is inappropriate. Compared to LDA, SSP could handle more complex situation where multiple subspaces co-exist in one semantic class. Besides, SSP can capture the structure of the subspaces better because geometric neighborhoods can be preserved while only semantic similarity is considered in LDA. Thus SSP could model the real scenario more accurately.

3.3 SSP with Relevance Feedback

In content-based image retrieval, relevance feedback is a powerful tool to enhance the learning of user's interest [18,19]. A general CBIR with Relevance Feedback works in the following way [20]:

(i) Given one or multiple query images from the user, a CBIR system searches the image database and ranks the images according to similarity to the query images.

(ii) The top *N* ranked images are represented to user for feedback.

(iii) The user gives semantic judgment on some of the retrieved images.

(iv) The CBIR system integrates the user's feedback into its algorithm and re-trains the algorithm to get more accurate results.

(v) The system goes back to step 2 until the user is satisfied or some other criteria are met.

In our semantic manifold learning method, the user feedback is used to refine the geometric and semantic relation graph between samples lying on single or multiple subspaces. Specifically if user gives relevance feedback on a small set of samples, we update Semantic Dissimilarity matrix as follows

$$SD^{(t+1)} = SD^{(t)} + RF^{(t)} \tag{27}$$

Where *t* is number of iteration and "+" denotes element-wise Boolean sum operation.

$$RF_{ij}^{(t)} = \begin{cases} 1 \text{ if } \boldsymbol{x}_{\text{i}}, \boldsymbol{x}_{\text{j}} \notin \text{same semantic class in user feedback} \\ 0 \text{ otherwise} \end{cases}$$

It is worth mention that our approach is different from I-LPP or ARE in that only semantic dissimilarity information is utilized from user feedback. The reason is that, as we indicate in Section 2, semantic dissimilarity information is more reliable to infer subspace structure because semantic similarity may arise from non-visual knowledge that can not be used in visual subspace discovery.

It is clear that the Geometric Similarity matrix *GeoSim* should remain the same. Thus the Geometric-Semantic Similarity matrix *GSSim* should be updated and normalized consequently as follows:

$$\begin{aligned}
GSSim^{(t+1)} &= \overline{SD^{(t+1)}} \cdot GeoSim \\
&= \overline{SD^{(t)} + RF^{(t)}} \cdot GeoSim \\
&= (\overline{SD^{(t)}} \cdot \overline{RF^{(t)}}) \cdot GeoSim \\
&= GSSim^{(t)} \cdot \overline{RF^{(t)}}
\end{aligned} \tag{28}$$

Since *GSSim* has to be normalized row-wise, we have

$$GSSim^{(t+1)} = GSSim^{(t+1)} / Diag(GSSim^{(t+1)}) \tag{29}$$

where the *Diag(A)* operation is defined as constructing a diagonal matrix where the *i*th diagonal element is sum of the corresponding *i*th row of the matrix *A*.

Note that in our method only semantic dissimilarity is adapted to update the Semantic-Geometric Graph while in the other two techniques, I-LPP and ARE, semantic similarity information is used to override the geometric relation between two samples. Consequently it is possible in their approaches that two images that are from different visual/geometric subspaces with the same semantic label in user feedback are forced to cluster together. That effort is unnecessary and may even deteriorate the classification performance.

The Geometric-Semantic Similarity Matrix ${S^{(t+1)}}_{Gs-Sim}$ can be efficiently updated in the following incremental way:

$$\begin{aligned}{S^{(t+1)}}_{GS-Sim} &= \sum_{i,j}(x_i - x_j)(x_i - x_j)^T {GSSim^{(t+1)}}_{ij} \\ &= (\sum_{i,j}(x_i - x_j)(x_i - x_j)^T ({GSSim^{(t)}}_{ij} \cdot \overline{{RF^{(t)}}_{ij}})) / Diag(GSSim^{(t+1)}) \\ &= ({S^{(t)}}_{GS-Sim} - \sum_{i,j}(x_i - x_j)(x_i - x_j)^T {RF^{(t)}}_{ij}) / Diag(GSSim^{(t+1)})\end{aligned} \tag{30}$$

Similarly the Dissimilarity Matrix S_{Diss} can be updated as below:

$${S^{(t+1)}}_{Diss} = {S^{(t)}}_{Diss} + \sum_{i,j}(m_i - m_j)(m_i - m_j)^T {RF^{(t)}}_{ij} \tag{31}$$

It is easy to find that $RF^{(t)}$ is a very sparse matrix because usually user can only give semantic judgment on very few images. From the above analysis, we can conclude that the user relevance feedback can be iteratively integrated into the subspace discovery with affordable computational expense.

4. EXPERIMENTS AND ANALYSIS

4.1 Test on Benchmark Datasets

The first experiment is designed to evaluate the performance of our proposed method on some benchmark datasets. Different setting of neighborhood size (K) is tested to find the optimal setting. The benchmark datasets tested are heart and breast-cancer (B.C.) dataset from UCI repository. The two data sets consist of pre-processed feature vectors of medical images. The two data sets have vector length of 13 and 9 respectively. In each run the program randomly pick up 170 and 200 for training and 100 and 77 for testing respectively. In all the experiments we run the program 100 times to get the average performance.

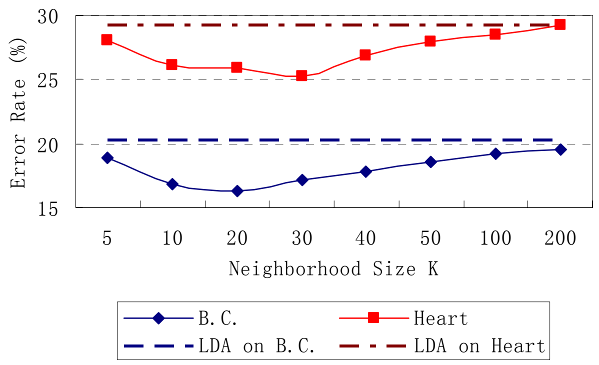
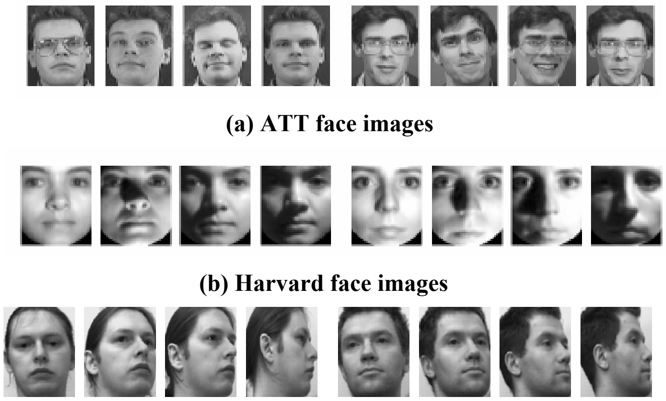


Figure 2. Test neighborhood size on Benchmark Dataset

The result in Figure 2 shows the effect of the size of neighborhood on the classification performance. It is clear that the size of the neighborhood does have influence on the performance of SSP. As the size of the defined neighborhood gets larger, according to our analysis in Section 3.2, SSP should reduce to well-known LDA. In Figure 2, the LDA performance is illustrated by two dotted lines. It is clearly that our proposed method performs better than LDA in most settings and will approach to LDA when the neighborhood size K gets larger and contains all the training samples. For simplicity, we use 20 as neighborhood size in the rest experiments since it gives the best performance in this one. However, it is worth mention that the optimal number of neighborhood size should be different with different data sets and is worthy further exploration.

4.2 Comparison to the State-of-the-Art

In the second experiment, we test the performance of our proposed method and compare it to the state-of-the-art techniques discussed in Section 2.3: LDA [4], LLE [6], S-LLE [15], LPP [9] and LDE [16]. ISOMAP is not tested because it's computationally expensive to calculate the global geodesic distance. K is set to 20 for our proposed method according to the previous experiment. The performance the state-of-the-art subspace learning techniques are optimized by tuning the parameters as suggested in their paper. For simplicity, nearest neighbor classifier is used in the projected space.



(a) ATT face images

(b) Harvard face images

(c) UMIST face images

Figure 3: Face images from three databases with different expressions (a), illumination conditions (b) and head poses (c)

We first apply those methods on face identification on three popular databases: Harvard [21], ATT [22] and UMIST [23] facial image databases. Harvard Face Image Database consists of grayscale images of 10 persons. Each person has totally 66 images which were classified into 10 sets. The Harvard database offers face images under dramatically different illumination condition. We use the set 1 as training set and choose set 1, 2 and 3 to test the performance as suggested in [21]. The ATT Face Image Database consists of 400 images for 10 persons. The images are from person with different expression and with or without classes. The grey images have resolution of 92×112 with different expressions, with or without glasses and almost same illumination condition. The UMIST Face Database consists of 564 images of 20 people, which covers a range of poses from profile to frontal views. We randomly assign one person's facial image as positive and the rest are considered as negative. 30 image features are extracted for each face images by PCA. In all three experiments two thirds of the images are randomly picked up as training sample and the rest is used for testing. The performance is evaluated by averaging accuracy over 100 runs, which is defined as the number of correctly retrieved face images over the number of all test images.

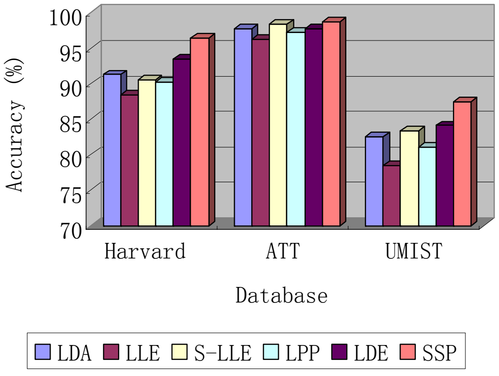
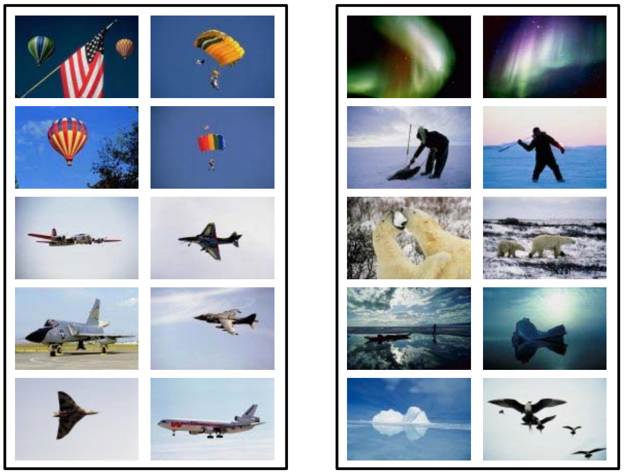


Figure 4: Comparison to the State-of-the-Art

In this experiment the facial images of different persons could be modeled by different subspaces. The task of telling the face images of a specific person from those of several other persons' constructs a multiple-to-one relation between the geometric subspaces and negative semantic class because all other person's face images are irrelevant to the query concept. From the results in Figure 4 we find that SSP performs best on all three databases.



(a) **(b)**

Figure 5: Pictures from two semantic classes in COREL database: (a) aviation and (b) north pole

4.3 Test on Image Classification

In the third experiment, we further test SSP and the state-of-the-art techniques on image classification application. The dataset used are COREL image databases. It contains 3000 color images which are roughly categorized into 30 classes. Each class contains 100 images. For simplicity in this experiment we randomly pick up two classes of images for classification. Two-thirds of the images are used for training while one third is used for testing. The image features we used are listed in table 1. In this experiment we only compare SSP with the supervised techniques.

Table 1: Image features used in the experiment

Feature Name	Description	Length
ColorHistNM	Normalized Color Histogram [24]	32
ColorMmtNM	Normalized Color Moments for HSV space [25]	9
WvNM	Normalized Wavelet Moments for texture [26]	10
WfNM	Normalized waterfilling feature for structure [27]	18

For comparing the image retrieval results, we use the *precision-recall* graph as the performance measure. *Recall* is a measure of the completeness of the retrieval sets, i.e., the percentage of retrieved objects in the correct answer. *Precision*, on the other hand, measures the purity of the retrieved set, i.e., the percentage of relevant objects among those retrieved.

From Figure 6 we can conclude that SSP outperforms other state-of-the-art techniques. Considering the rich content of images in COREL database, it could be explained by that our method not only preserves the subspaces within same semantic class but also separate the subspaces with different semantic labels, which can facilitate classification.

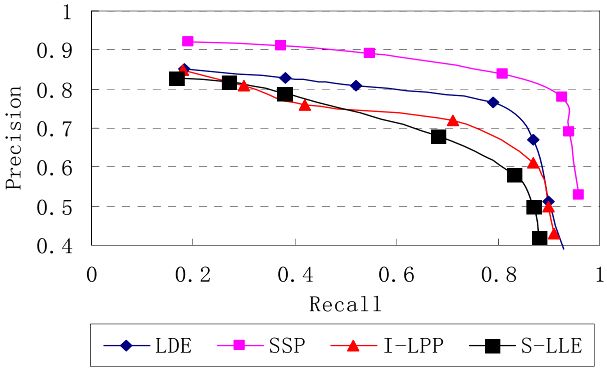


Figure 6. Precision-Recall graph for test on COREL database

4.4 Test SSP with Relevance Feedback

In the fourth experiment we test SSP with Relevance Feedback on image retrieval application. The data set is the same COREL data set as in Section 4.3. In this experiment we randomly pick up one image as query and the program ranks the images according to their similarity to the query image. Top ten most similar ones are presented to user for feedback. Note that for simplicity purpose our test uses an automatic pseudo relevance feedback scheme, where machine could give relevance feedback to the top 10 retrieved images based on the ground truth information. In this experiment, the precision is defined as number of correct images in top 50 retrieved images. The reported result is the average over 100 runs of the experiment.

Figure 7 shows the precision when the data are projected to different dimension *D* and relevance feedback are integrated iteratively. It is clear that SSP improves the performance by learning more semantic information from user over iteration. It is also worth mention that SSP could project the data to larger than *C-1* dimension (*C=2* in our two-class case) and obtained more accurate classification in those spaces. Compared to traditional technique, such as LDA, it is favorable because complex data may be better modeled in larger than *C-1* dimension space, e.g. *D=20* in this experiment.

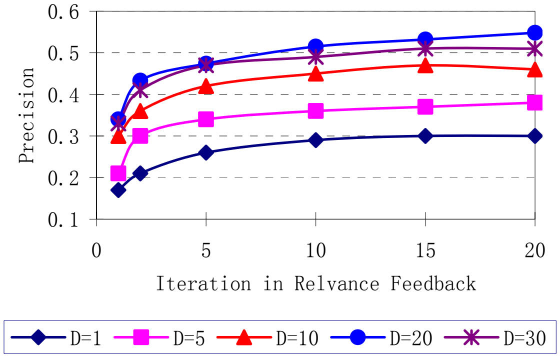


Figure 7. Performance over iteration on different projection dimensions

Figure 8 shows the comparison of SSP with Relevance Feedback with S-LLE, I-LPP and ARE. Note that S-LLE is not a semi-supervised technique. Although S-LLE also benefits from the user feedback because more ground truth information about images is known to the classifier. It is no wonder to find that semi-supervised techniques I-LPP, ARE and SSP+RF improves more from feedback than S-LLE. Compared to I-LPP and ARE our technique works better because semantic dissimilarity is better utilized in discovering subspace while the other two methods try to use semantic similarity to override geometric neighborhood information.

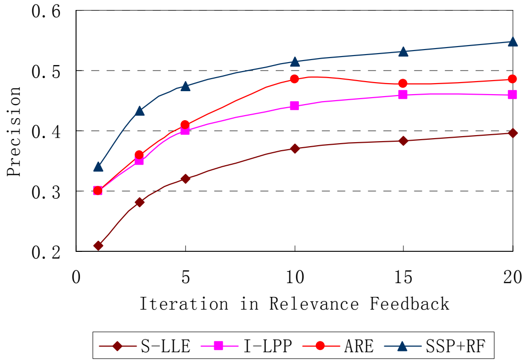


Figure 8. Comparison with other semi-supervised methods

5. CONCLUSIONS AND FUTURE WORK

In this paper we propose a novel framework SSP that learns the mapping between semantic concept and geometric subspaces for image retrieval and classification tasks. Theoretical analysis proves that our proposed method reduces to well-known LDA when one semantic class corresponds to one geometric subspace. Relevance feedback can be interactively integrated into our framework to refine the semantic concept learning.

Due to the rich visual content of images, SSP models the images as lying on subspaces embedded in the geometric feature space. It is more appropriate for content-based image retrieval than the traditional approaches, such as PCA and LDA, which often assume all image data are from a specific distribution model. Unlike most existing subspace learning techniques, SSP aims at not only preserving the subspace structure in the projected space but also separating subspaces from different semantic classes in the projected space, which should not be neglected for classification. Our further study on the relation between semantic class and geometric subspaces indicates that user's semantic interest may arise from non-visual knowledge and consequently one semantic class may contain multiple image subspaces which have different geometric structures. That problem has not been addressed or properly handled in existing subspace learning techniques. Based on this analysis, SSP utilizes both semantic and geometric similarity to define local neighborhoods while semantic dissimilarity is used to separate different semantic classes. To further enhance the retrieval accuracy, user relevance feedback could be incrementally integrated in SSP for capturing the semantic concept dynamically. Experiments are designed and conducted on benchmark datasets and two image retrieval applications for performance evaluation. The results have shown the superior performance of our proposed method.

This research work will be continued in the following directions: 1) instead of using feature vectors as images data, 2-D matrix representation may enhance the performance of SSP; 2) kernel transformation may be used to better separate non-linear subspaces; 3) local neighborhood into meaningful may be clustered in a self-discovery way.

Acknowledgement: This work was supported in part by the Army Research Office (ARO) grant under W911NF-05-1-0404, and by the Center of Infrastructure Assurance and Security (CIAS), the University of Texas at San Antonio. We thank the anonymous reviewers for their review comments and John R. Kender for the guidance in revising the paper.

6. REFERENCES

[1] A. W. M. Smeulders *et al.*, "Content-based image retrieval at the end of the early years," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349 - 1380, 2000.

[2] Q. Tian, J. Yu, Q. Xue, N. Sebe, "A New Analysis of the Value of Unlabeled Data in Semi-Supervised Learning for Image Retrieval," *Proceedings of International Conference on Multimedia and Exposition*, Taipei, Taiwan, ROC, June 27-30, 2004.

[3] I. T. Jolliffe, *Principal Component Analysis*, 2nd edition, New-York: Springer-Verlag, 2002.

[4] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc., 2001.

[5] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831-836, 1996.

[6] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, 290:2323-2326, December 2000.

[7] J. B. Tenenbaum, V. de Silva and J. C. Langford. "A global geometric framework for nonlinear dimensionality reduction," *Science*, 290: 2319-2323, December 2000.

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

[9] X. He and P. Niyogi, "Locality preserving projections," *Neural Information Processing Systems 16 (NIPS'2003)*.

[10] X. He, "Incremental semi-supervised subspace learning for image retrieval," *Proc. of ACM Multimedia*, 2004.

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

[12] M. Zhu and A. Martinez, "Optimal subspace discovery for discriminant analysis," *Workshop on learning in CVPR*, 2005.

[13] M. Figueiredo and A. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. on PAMI*, 24(3):381-396, 2002.

[14] R. Duin, E. Pekalska, P. Paclik, and D. Tax, "The dissimilarity representation, a basis for domain based pattern recognition?", *ICPR 2004 Workshop Proceedings*, Cambridge UK, 22 August 2004.

[15] D. de Ritter *et al*. "Supervised locally linear embedding," *Proc. of ICANN/ICONIP*, 2003

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

[17] Y. Lin, T. Liu and H. Chen, "Semantic manifold learning for image retrieval," *Proc. of ACM Multimedia*, August 2005

[18] 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.

[19] Y. Rui and T. Huang, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. on Circuits and Video Tech., Special Issue on Segmentation Description, and Retrieval of Video Content*, vol. 8, 1998.

[20] S. X. Zhou and Thomas S. Huang. "Relevance feedback for image retrieval: a comprehensive review," *Multimedia Systems,* 8(6):536--544, 2003.

[21] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711-720, 1997.

[22] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision,* Sarasota FL, December 1994.

[23] D. Graham and N. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," *Face Recognition: From Theory to Applications,* NATO ASI Series F, Computer and Systems Sciences, Vol. 163, pp.446-456, 1998.

[24] M. Swain and D. Ballard, "Color indexing," *Intl. Journal Computer Vision*, vol. 7, no.1, pp. 11-32, 1991.

[25] M. Stricker and M. Orengo, "Similarity of color images," *Proc. SPIE Storage and Retrieval for Image and Video Databases*, pp. 381-392. 1995.

[26] J. R. Smith and S. F. Chang, "Transform features for texture classification and discrimination in large image database", *IEEE Intl. Conf. on Image Proc*., 1994

[27] S. X. Zhou, Y. Rui, and T. S. Huang, "Water-filling algorithm: a novel way for image feature extraction based on edge maps," *IEEE Intl. Conf. on Image Proc.*, 1999.

[28] S. X. Zhou and Thomas S. Huang, "Small sample learning during multimedia retrieval using BiasMap," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.