Cross-Media Manifold Learning for Image Retrieval & Annotation

Xianming Liu, Rongrong Ji, Hongxun Yao, Pengfei Xu, Xiaoshuai Sun, Tianqiang Liu
School of Computer Science and Technology, Harbin Institute of Technology
No.92, West Dazhi Street, Harbin, P. R. China, 150001
Mail: {liuxianming, rrji, yhx, pfxu, xssun, tqliu}@vilab.hit.edu.cn Tel: +86+45186416485

ABSTRACT
Fusion of visual content with textual information is an effective way for both content-based and keyword-based image retrieval. However, the performance of visual & textual fusion is affected greatly by the data noise and redundancy in both text (such as surrounding text in HTML pages) and visual (such as intra-class diversity) aspects. This paper presents a manifold-based cross-media optimization scheme to achieve visual & textual fusion within a unified framework. Cross-Media manifold co-training mechanism between Keyword-based Metric Space and Vision-Based Metric Space is proposed creatively to infer a best dual-space fusion by minimizing manifold-based visual & textual energy criterion. We present the Isomorphic Manifold Learning to map the annotation affection in image visual space onto keyword semantic space by manifold shrinkage. We also demonstrate its correctness and convergence from mathematical perspective. The retrieval can be performed using both keyword or sample images respectively on Keyword-based Metric Space and Vision-Based Metric Space, while the simple distance classifiers will satisfy. Two groups of experiments are conducted: The first group is carried on Corel 5000 image database to validate our effectiveness by comparing with state-of-the-art Generalized Manifold Ranking Based Image Retrieval and SVM. The second group is done over real-world Flickr dataset with over 6,000 images to testify our effectiveness in real-world application. The promising results show that our model attains a significant improvement over state-of-the-art algorithms.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Retrieval Models.
General Terms
Algorithms, Measurement, Experimentation

Keywords
Automatic image annotation, Manifold learning, Content-based image retrieval, Web image search, Co-training

1. INTRODUCTION
The explosive growth of Internet facilitates easy access of gigantic image volume in our daily lives. To search and classify such gigantic image collections poses significant technical challenges. Content-based image retrieval is a feasible solution to address this issue. However, the performance of CBIR is limited because of the semantic gap though numerous efforts made over the past decade. To boost the retrieval performance, it has been a common scene to fuse content-based and keyword-based solutions to reinforce each other in effective image search. The topic of visual & textual fusion for multimedia retrieval is becoming research hot spot in recent years [17, 22].

The task of visual & textual fusion can be described as the combination of multimedia evidence from different aspects to achieve best performance in retrieval. For retrieval, fusion of visual & textual content usually refers to “Multimodel Image Retrieval”, in which many methods have been proposed for this issue. X. Wang [2] divides the former approaches into three groups: 1) Combining different models using simple linear combinations. 2) Restoring to human interaction. 3) Probabilistic models. However, there are still mainly two drawbacks existing in these approaches among the three groups. Firstly, the results of retrieval rely on the learning result, which will be affected greatly by the distribution of the training samples in the feature space. Secondly, if the training sample is labeled wrongly because of the noise, the learning results remain inaccurate during the propagation procedure.

Recently, graph-based approaches have been used in image annotation and retrieval [6]. These methods propagate the keywords along the vertex of graph for image annotation. The keywords may come from different sources, such as manually labeling or automatic extraction results from the surrounding text of web images into image classifier learning [2, 6]. However, it suffers from small sample learning problem due to the complexity to approximate the graph structure from original feature space to real semantic space [2].

In many cases, the surrounding text extracted from the web image contains lots of irrelevant noise. For instance, an advertised image embedded in a news report is prevalent in news websites. It would not be feasible to treat them equally confident as the visual features. Besides, there is redundant information in such text description. Another property of the web images and texts is the asymmetry issue. Huge amount of data from the web is not structural since: 1) the data is not well labeled 2) the data set is open which means it can’t be clustered.

On the other hand, the semantic revealing ability of solely visual features is limited, which calls for co-training strategy between visual features and textual descriptors, usually including semi-supervised learning and transductive learning [8-11].

Manifold ranking is adopted in CBIR and annotation [18-20, 22] as a kind of transductive learning, by modeling the original image database as the “image space”, and involving the users’ relevance feedback to “smooth” this manifold to reveal human’s image perception. Its merits lie in its perceptual simulation ability and excellent computational efficiency.
This paper considers the retrieval task between two models as manifold learning scenario, based on which a learning model can be expressed as a manifold mapping between visual & textual feature spaces. From manifold learning aspects, visual & textual space can be associated using Isomap, based on which either image annotation or keyword-based retrieval can be achieved. In manifold learning, the change of one sample will lead to the deformation of the manifold. So a mapping which can preserve the consistency between different models is needed to complete the deforming procedure. We define such a manifold with this mapping as Isomorphic Manifold. To address data asymmetry issue mentioned above, we train our Isomorphic Manifold using labeled data and their underlying relationship with unlabeled data, as the common agreement of its semi-supervised nature [8-11].

The rest of this paper is organized as follows: Section 2 introduces the Cross-Media Manifold Learning strategy for Image Retrieval and annotation. To solve the problem of mapping the movement in low level dimensional space back onto the high level, this part also involves the Isomorphic Manifold Learning. Its correctness and convergence proof are given in Section 3. The experiment is introduced in Section 4 while the experimental results are shown in Section 5. At the end of this paper, we make a conclusion.

2. CROSS-MEDIA MANIFOLD LEARNING FOR IMAGE RETRIEVAL

2.1 Cross-Media Manifold Learning for Image Retrieval

As addressed in the first section, it is reasonable to fuse the text information (such as surrounding text of image embedded in HTML page) and its visual information (such as the color and texture features). The description from HTML pages illustrates the image sufficiently (although including much redundancy) and the visual information is regarded as reliable enough. Based on this assumption, we can correct the annotation labeled by keywords with the visual information, to achieve textual noise elimination. We propose a novel Cross-Media Manifold Learning.

The basic idea is to treat the distribution of keyword annotation results as the original manifold and perform the manifold reduction to get a graph G. To eliminate the redundancy and noise existed in the initial manifold the visual information is involved in the training stage on G. The distributions of these two different features can be formulated as metric spaces as:

**Keyword-Based Metric Space:** The images can be labeled with the keywords extracted from corresponding HTML by performing the TF-IDF rule in Information Retrieval [24]. All the keywords can expand an orthogonal space by removing the synonyms. We call this orthogonal space as a Keyword-Based Metric Space, and the value on each dimension is the probability that the image can be labeled with a specific keyword. We denote the Keyword-Based Metric Space as the matrix \( K \), and each image as a vector from a given row of \( K \) which is denoted as \( K_i \) for image \( i \). If the total number of dimensions is \( N \), the vector can be written as \([K_{i1}, K_{i2}, ..., K_{in}]\) for image \( i \). The Cosine Similarity on Keyword-Based Metric Space is defined as:

\[
\text{Sim}_{ij} = \frac{\sum_{i=1}^{N} (K_{ij} \cdot K_{kj})}{\sqrt{\sum_{i=1}^{N} K_{ij}^2 \cdot \sum_{i=1}^{N} K_{kj}^2}}
\]

(1)

Which is bounded in (0, 1), and a small value indicates a low similarity while a large value corresponds to a high similarity.

**Vision-Based Metric Space:** It is the common visual feature space in Content-Based Image Retrieval which is constructed with the visual features such as color and textual. Corresponding to the Keyword-Based Metric Space, the distance adopts the L2 distance among the Vision-Based Metric Space.

The metrics on these two different spaces are Keyword-Based Metric and Vision-Based Metric separately. We can construct the cross-media manifold based on the two metric spaces by using Manifold Reduction such as Isomap [3] on Keyword-Based Metric Space to get the graph \( G \) with the weights of each edge standing for the similarity between two vertices. \( G \) can represent the underlying relationship within all the samples. To correct the manifold with noise, the L2 distance defined in Vision-Based Metric Space is involved. It is still a big problem to map the movement of data samples from low dimensional space back to the original high dimensional space [25]. We propose the Manifold Isomorphic Learning to address this problem well.

2.2 Manifold Isomorphic Learning

2.2.1 Manifold Reduction & Learning

To map the changes in low-dimensional space back onto the high-dimensional space, Manifold Reduction and Manifold Learning are formulated by mathematicians in this section. Definition 1 and Definition 2 are defined on this purpose.

**Definition 1:** Two metric spaces \((M_1, d_1)\) and \((M_2, d_2)\) are said to be topological homeomorphic if there exists a homeomorphism (continuous maps with continuous inverses) between the two spaces.

**Definition 2:** A manifold can be expressed in the form of a triple-set:

\[ \{R^d, R^n, f\}, d < m \]

\( R^d \) is the d-dimensional metric space the manifold embedded in and \( R^n \) is the m-dimensional metric space in which all the initial points distribute while \( f \) is a homeomorphism defined as:

\[ f: C \subset R^d \rightarrow R^n \]

which can retain the topological structure in the space \( R^n \), with \( C \) a compact subset of \( R^d \). Based on the definitions above, we can distinguish the Manifold Reduction and Manifold Learning as: Manifold Reduction aims to find a subspace \( C \) while Manifold Learning aims to simulate the homeomorphism mapping \( f \) of metric space \( C \). Because the procedure of simulating the homeomorphism relies on the distribution of metric space \( C \), Manifold Reduction acts as the basis of Manifold Learning. There have been a lot of approaches used for Manifold Reduction [3] [15].

It has been proved that an undirected graph \( G \) can be turned into a metric space [13], consequently it is reasonable to replace the compact subset \( C \) with graph \( G \). As a result, the transformation from a high dimensional Euclidean Space into a Graph satisfies the definition of Manifold Reduction, which is the state-of-the-art method in Manifold Reduction.

On the other hand, the target of Manifold Learning is to describe the mapping. Authors in [12] proposed a method to minimize the error that occurs in the simulating which is defined in a linear form. In this paper, we will introduce an Isomorphic mapping to simulate the homeomorphism by defining Isomorphic Manifold.
2.2.2 Manifold Isomorphic Learning

To achieve the target in Section 2.1, we formulate the Keyword-Based Metric Space as the original high dimensional space while the Vision-Based Metric Space as the low dimensional space in real-world application for our Isomorphic Manifold Learning. They are marked as M and C respectively. The homeomorphism is defined as

\[ f \in M \rightarrow C \]  \hspace{1cm} (2)

We define an isomorphic mapping function \( g \) in form of

\[ s(x', y') = g(f(x), f(y)) = h(d(x, y)) \]  \hspace{1cm} (3)

Where \( x \) and \( y \) are two points on behalf of sample images in \( M \), \( x' \) and \( y' \) the corresponding points in \( C \); \( d \) and \( s \) are the metrics defined in \( M \) and \( C \) separately, in the forms of Keyword-Based Metric and Vision-Based Metric; \( h \) is a continuous and monotonous function named Isomorphic Function. As proven in [13], \( h \) is reversible and can preserve the metric.

Since the main target is to fuse the two different models together and as a result to eliminate the noise, co-training is adopted as the strategy. The co-training deals well with the condition that the features can be separated into two disjoint sets [26]. Thus, the co-training can be performed on the manifold from two different metrics: Keyword-Based Metric & Vision-Based Metric defined in section 2.1. The algorithm firstly performs the manifold reduction using Isomap [3] to map the Keyword-Based Metric Space onto the Vision-Based Metric Space. Since the noise and redundancy existed in the text descriptor, the reduction results are filled with errors. To eliminate the errors and reduce the semantic gap we perform co-training between \( M \) and \( C \) as follows:

1. Manifold is initialized in the Keyword-Based Metric Space \( M \). The distribution of all sample points reflects the initially condition of text descriptors.
2. Performing the ISOMAP [3] to map the distribution of \( M \) onto the Vision-Based Metric Space \( C \). The manifold reduction result is represented as a Graph \( G \).
3. The similarities defined in equation (1) between each pair of adjacent vertices are formulated as the weights of edges on \( G \).
4. Adjust the weights of graph \( G \) comparing the Visual-Based Metric between any two adjacent pair of points using the Isomorphic Function which will be illustrated in Section 2.3.1.
5. Performing the Manifold Shrinkage strategy to map the changes on \( G \) back onto \( M \) by modifying the weights of corresponding keywords in vector .
6. If the iteration doesn’t converge, go to step 2. Else stop iteration and go to step 7 to make retrieval.
7. Retrieval on the trained manifold.

The “Shrinkage” and “Isomorphic Function” will be shown in section 2.3, while the proof of convergence of shrinkage in section 3. The convergence criterion can be defined as:

**Averaged Convergence Criterion:** The convergence is defined based on the Averaged Iteration Difference defined as:

\[ \frac{\sum |s^{t} - s'|}{\text{imgCount}} \]  \hspace{1cm} (4)

Where \( \varepsilon > 0 \) and \( \text{imgCount} \) is the total number of the images in database.

To retrieval on the Cross-Media Manifold for Image Retrieval, it comes mainly two ways: 1) retrieval on Manifold \( M \); & 2) retrieval on Graph \( G \). For CBIR, the keyword-based query can be performed based on approach 1) while the content-based query performed based on 2). As for the classification task, the simple distance classifier is enough. To query on \( M \), \( L2 \) distance is available while on Graph \( G \), the Graph-based distance is better.

2.3 Isomorphic Function & Manifold Shrinkage

Since most manifold reduction methods are nonlinear, the inverses of homeomorphism don’t exist or can’t be represented. Therefore, the change in low-dimensional space can’t be reflected in the high-dimensional space directly.

The shrinkage of manifold is proposed to solve this problem and aims at formulating the change between the metric spaces \( M \) and \( C \). The movements of vertices in the low-dimensional metric space (Graph \( G \)) are mapped back onto the high-dimensional one (Manifold \( M \)) by performing the shrinkage based on the Isomorphic Function \( g \) which is defined in equation (3).

2.3.1 Isomorphic Function

The Isomorphic Function in form of iteration function should satisfy the following criterions:

1. **Differentiable and continuous:** Since the manifold should be differentiable while a differentiable function must be continuous.
2. **Reversible:** This property is based on the definition of Isomorphic Manifold.
3. **Convergent:** Keep the iteration stop at the limit.

Thus, we define a type of Isomorphic function in iteration form as

\[ h(x, y) = s^{t+1}(x, y) = \exp\left(-\frac{d(x, y)^2}{\mu \cdot s'(x, y)}\right) \]  \hspace{1cm} (5)

Where \( s'(x, y) \) is the the weight of edge between the vertices \( x \) and \( y \) at the \( t \)th round of the iteration on graph \( G \) reflecting the similarity between these two sample images from the perspective of keywords. And \( d(x, y) \) represents the metric defined on the \( C \), which can be defined as the \( L2 \) distance between these two images from the perspective of visual content. A larger similarity \( s \) indicates small distance which represents that the two samples are more similar in current metric space. \( \mu \) is a const to adjust the speed of the iteration and normalization between two metrics from different spaces. Along all the iterations, the similarity \( s \) is bounded between 0 and 1 by the exponent function since \(-\frac{d(x, y)^2}{\mu \cdot s'(x, y)}\) is always negative.

2.3.2 Shrinkage of Manifold

Shrinkage of Manifold mainly deals with the problem of mapping the movement in \( G \) back into \( M \). In an intuitive sense, this procedure forces the points in \( M \) making the displacement along with the movement of the corresponding point in on graph \( G \). In our application, it indicates that the position of points in Keyword-Based Metric Space \( M \) changes with the same rate as the movements on \( G \) compared to the Vision-Based Metric Space \( C \). As a result of the movement in \( M \), the weight of each keyword to
annotate the given image is modified referring to the visual similarities. To achieve this goal, we firstly define a shrinkage rate based on the \((t+1)\)th iteration as:

\[
\rho'(i, j) = \exp\left(\frac{s^{t+1}(x, y)}{s(x, y)} - 1\right) - 1
\]  

(6)

It is obvious that:

If \(s^{t+1}(i, j)\) is larger than \(s(i, j)\), which indicates the two images are more similar, the shrinkage rate \(\rho(i, j)\) is positive;

If \(s^{t+1}(i, j)\) is smaller than \(s(i, j)\), which indicates the two images are less similar, the shrinkage rate \(\rho(i, j)\) is negative;

Suppose the vector \(K\) represents the position of point \(i\) lying in metric space. Then the shrinkage can be formulated as:

\[
K_t(t) = \left[K_t(t-1) + \sum_{r=1}^{\infty} (K_t(t-1) - K_t(t-1))\rho'(i, j)\right] \quad t > 1
\]

(7)

\[
K_t(t) = 0
\]

(8)

Where \(X\) is the set of points adjacent with point \(i\) in the graph \(G\) obtained by manifold reduction.

For each pair of adjacent points \(i\) and \(j\) in round \(t\), the distance in the metric space where the manifold embedded in should also be adjusted by the same rate: when the similarity becomes larger, the distance on manifold should be shortened by the rate of \(\rho(i, j)\); when the similarity becomes smaller, the shrinkage rate \(\rho(i, j)\) is positive.

By Performing the Manifold Shrinkage strategy for each point in each-round of iterations, our Cross-Media Manifold could be trained to fit for the distribution of all the points referring to both the keywords and visual content distributions. Thus, we can get a shrunk metric space which reflects the intrinsic relationships among all image sample points.

3. Convergent: Since from Equation (8) we can obtain:

\[
\frac{\partial h}{\partial s} = \exp\left(-\frac{d^2}{\mu s}\right) \left(-\frac{d^2}{\mu s} - \frac{2d^2}{\mu^2 s^2}\right) > 0
\]

It indicates that, if the similarity of the results from the manifold reduction plays a more important role in the iteration, the similarity will increase in the iteration stage, which means the description of the manifold is not sufficient.

From Equation (9), we can obtain

\[
\frac{\partial h}{\partial d} = \exp\left(-\frac{d^2}{\mu s}\right) - \frac{2d}{\mu^2 s^2} < 0
\]

If the distance of metric \(d\) in the graph metric space takes a more important role, the similarity will decrease when iterations increase since the manifold describes two points closer than they distribute underlying. When the iteration stops, it will come to an utmost that balances both metric \(d\) and similarity \(s\).

4. EXPERIMENTAL SETUP

4.1 Database Selection

Two image databases are adopted in performance evaluation:

Corel 5000 Image Database: It contains 50 classes with each class 100 images. 20 images from each class are manually labeled. The total number of the keywords is 202. The initial values in vectors for any image are either 0 or 1: if the image is manually labeled with keyword \(j\), then \(K_{ij}\) is 1; otherwise \(K_{ij}\) is 0.

To avoid the zero vectors of unlabeled samples in training stage, we involve a random keyword for an unlabeled image out of the 202 keywords. This also brings noise, but would be shown eliminated after training.

Flickr Image Database: This database contains 6000 images downloaded from www.flickr.com together with its textual descriptors in HTML page. The keywords are extracted using the TF-IDF rule. The element \(K_{ij}\) in Section 2.1 is initialized by the probability of keyword \(j\) labeling image \(i\), which can be calculated by counting the frequency of keyword \(j\) contained in the document image \(i\) embedded in.

4.2 Feature Extraction

As for Vision-Based Metric Space, we use 360-dimension Color histogram features in H component of HSI color space and 8-dimension texture co-occurrence features [7] both for Corel Database and Flickr Database. We adopt L2 distance as the similarity metric in our experiments.

4.3 Model Parameter Setting

The algorithm for Manifold Reduction adopts the K nearest neighbor based ISOMAP method [3]. And the number of nearest neighbor \(K\) is set as 50 according to the volume of the database. The degree of nearness can be explained as: the larger similarity \(Sim_{ij}\) between points \(i\) and \(j\) indicates the smaller distance between them.

4.4 Evaluation Criterion

Precision-Recall curve is a state-of-the-art method to evaluate the performance of the image classification system. But for image retrieval system, the users pay more attention on the precision than the recall, especially for search engines. Besides, the retrieval system should emphasize more on how to make the top ranks more accurate as well. From the user’s point of view, a more
satisfying result at the top ranks of the returned list will be more valuable. Therefore, besides precision measurement, we also adopt the NDCG (Normalized Discounted Cumulative Gain) [23] as one of our performance evaluation measures which is based on the following assumptions:

1. Highly relevant documents & images are more valuable.
2. The greater the ranked position of a relevant document & image, the less valuable it is for user.

NDCG is defined as:

$$N_i = n \sum_{j=1}^{m} \frac{r(j)}{\log(1 + j)}$$

Where $N_i$ is the NDCG at $i$’s rank position, and $m$ is the last position of the relevant sample within the first $i$ samples in the ranked list. $r(j)$ represents the rate of relevant between the $j^{th}$ sample in the ranked list and the query bounded in $[0,1]$. $n_i$ is the normalized const for a given ranked list.

This NDCG evaluation measure has the advantages as follows:
1. Graded: it is more precise than P-R
2. Reflect more user behavior (e.g. user persistence)
3. Sensitive to position of the highest rated page.
4. Normalized for different lengths lists.

In our experiments, we adopt the following measures to evaluate the performances of Isomorphic Manifold Learning: 1) Precision of a given query. 2) The average precision for all queries. 3) NDCG@1, NDCG@3, NDCG@5.

4.5 Query Scenario

The user query modes can be mainly divided into two categories: query by keywords and by sample images. For web search engine and other retrieval tasks, the former one takes main part, such as http://images.google.com and http://image.yahoo.com. From the users’ point of view, to find or browse the images of a given topic in the form of keyword is the consensus when using retrieval system. In our experiments, we adopt the keyword-based query as the default mode.

Besides, we also perform the retrieval task queried by sample images in our experiments in order to show the influence of textual descriptors to visual content. Since most of Manifold Learning mechanisms are adopted in form of querying by sample images, it is used for performance comparison with the state-of-the-art methods as well. As mentioned in Section 2.2.2, the simple distance classifiers are adopted. The distance metrics for the two spaces are L2 distance and the shortest path distance on graph separately.

5. Experimental Results

The experiments are performed separately on the two databases described in Section 4.1: Corel Database and Flickr Database. The experiment performed on Corel is mainly used to validate the correctness of Isomorphic Manifold Learning and the experiment performed on Flickr aims to testify our performance in real-world scenario.

5.1 Algorithm Validation & Performance Comparison

To validate the correctness of our algorithm and evaluate the performance, the experiment is firstly performed on Corel 5000 database mentioned in section 4.1. We randomly choose 20 keywords from the keyword list to make queries. Besides, as mentioned in Section 2.2.2, we perform the experiment queried by sample images as well, in order to compare our method to other approaches. Figure 1 shows the top 50 samples in the retrieval results queried by keywords “bus” on Corel.

To illustrate the convergence of Isomorphic Manifold Learning iteration, the Averaged Iteration Difference (AID) defined in section 2.2.2 is recorded and shown in Figure 2. The abscissa axis is the first 50 round of iteration while the ordinate is the AID for each round. It is obvious that the algorithm satisfies the convergence condition mentioned in section 2.3.1.

5.1 Algorithm Validation & Performance Comparison

To validate the correctness of our algorithm and evaluate the performance, the experiment is firstly performed on Corel 5000 database mentioned in section 4.1. We randomly choose 20 keywords from the keyword list to make queries. Besides, as mentioned in Section 2.2.2, we perform the experiment queried by sample images as well, in order to compare our method to other approaches. Figure 1 shows the top 50 samples in the retrieval results queried by keywords “bus” on Corel.
precisions are still higher than 0.5, which indicates that our algorithm can propagate the annotation results well.

Figure 4: Averaged Precision comparison between our method, Generalized Manifold Ranking Based Image Retrieval [19] and only using SVM.

Figure 4 shows the comparison results of averaged precision between our method, Generalized Manifold Ranking Based Image Retrieval (GMRBIR) proposed in [19] and only using SVM on the same database querying by sample images. Since there are few methods suitable to be compared with by querying in keywords, we here show the improvement in performance by querying in samples images. Meanwhile the red line shows the performance of querying by keywords. Obviously, there is a significant boost for performance compared with the GMRBIR as well as SVM.

Figure 5: NDCG@1, NDCG@3 and NDCG@5 for 10 randomly given queries.

Figure 5 shows the NDCG measure evaluation for the given 10 query keywords selected from the 20 ones mentioned above. It indicates that all of the first returned samples are relevant to the query, and all of the results within the top 3 returned samples are relevant except the one for keyword “water”, and it becomes only 2 irrelevant for the top 5 samples.

5.2 Experiment on Web Images Database

This experiment aims to illustrate the solution for image search in web-based real-world application.

This part of experiment adopts the Flickr Database containing 6000 images, and totally 4287 keywords in the keyword list extracted from the text descriptors. Within the descriptions of images, there exist much noise as well as redundancy because that they are not manually modified and limited.

Figure 6 shows the convergence condition of the experiment performed on Flickr Database. Compared with Figure 2, the curve is not as smooth as the one in Figure 2. This is because the noise affection in the text descriptors.

Figure 7: Averaged Precision curve compared with the precision curves of some keywords for Flickr Database.

Figure 7 shows the average precision for randomly chosen 10 keywords out of all the 4287. The black line shows the average precision of the 10 keywords returning 20, 25, 30, 35 and 40 images respectively. And the broken line in black is the keyword with best performance (“cloud”). The most interesting one is the line in red, which is the precision curve of the keyword “natura”, a wrong written keyword existed in the keyword list. It’s similar with the one of “nature” as shown in Figure 8. This issue will be further discussed in Section 5.3.

Figure 8: The positive samples within top 50 returning images which “nature” and the wrong written “natura” shares.

Figure 9 shows the NDCG evaluation measures for the Flickr Database. Though at the 1st position the query results could not
exactly satisfy the users’ demand, most of the correct results can do within the top 5.

![Figure 9: NDCG@1, NDCG@3 and NDCG@5 for 10 randomly given queries on Flickr Database.](image)

5.3 Further Discussion

5.3.1 Visualization of Manifold

Figure 10 shows the visualization of Keyword-Based Metric Space using PCA on Flickr Database before and after training in (a) and (b) respectively. It is obvious that the ground truth in (a) has three main problems compared with the results of Isomorphic Manifold Learning shown in (b):

1. The labels for images are too specific but not generic, which will lead to Over Fitting during the classifying process.
2. The keywords or semantic items are not propagated to other image samples. They are clustered together around each semantic item.
3. The relationships between each pair of semantic items or keywords are cut off. As shown in (b) of Figure 10, our method deals well with this problem and approaches this relationship by using the visual similarities between images.

![Figure 10: Visualization of the Keyword-Based Metric Space using PCA (Flickr Database).](image)

5.3.2 Semantic Similarity

Considering the 3rd problem in Section 5.3.1 and our experiment results, the relationships between semantic items should be involved, such as synonym, the case of pluralism and the case of wrong-written. A lot of methods have been proposed to solve this problem, such as the WordNet. WordNet [27] provides a way to link annotated images together. It assumes that if we can find one relevant image in an additional annotated source, then we can locate other semantically similar images via the WordNet links [17]. However, WordNet is initially used for the purpose of textual analysis, the goal of which is diverse from the one of image annotation and CBIR. On the other hand, the methods such as WordNet can not reveal the correlation of images from the content-based perspective. In annotation keyword pruning, the pruning rule in WordNet is based on solely linguistic association of words, which is fixed and ignores the ground-truth visual content of annotated image. For instance, when user is querying for concept “apple”, even when he/she gets tens of feedback images, the correlations of “apple”-“computer” and “apple”-“fruit” still cannot be adjusted in WordNet with assistant of image visual content. Therefore it is suitable to evaluate the relationship between content-based keywords from the perspective of visual-content based similarities. Similar with the cognitive process of human beings, the similarities of keywords should be judged by their expressive force in representing the visual content. For instance, during the keyword annotation procedure, the keywords “sea” and “ocean” are synonyms because their results of annotation are nearly the same. To evaluate the semantic relationships among different keywords in keyword list following this rule, the cosine similarities are calculated with all the keywords in both Corel 5000 & Flickr Databases. Table 2 shows part of the evaluation results. It is obvious that the similar keywords such as “wood” and “trees” will have the nearly same query results. Meanwhile, the Cross-Media Manifold and Isomorphic Manifold Learning are tolerant to pluralism and wrong written keywords. It is also sensitive to the keywords which are semantic relevant such as “green” vs. “colors”, “France” vs. “Europe” and “flower” vs. “leaf”. As the result, the similar keywords can be propagated in a similar way, causing the similar annotation results. Therefore, the Isomorphic Manifold Learning can deal well with the problem of synonyms on itself.

![Table 2: Keywords evaluation results for Flickr Database](image)

6. CONCLUSION AND FUTURE WORKS

In this paper, we exploit the manifold-based fusion of visual content and textual descriptors from cross-media manifold-based
optimized learning aspect, and consequently construct a
generalized and effective image retrieval & annotation framework. Our cross-media manifold learning: Isomorphic Manifold Learning is with mathematical correctness proven as well as superior performance comparing with state-of-the-art algorithms. We design experiments on Corel to testify our performance with baseline comparison, while we also design experiments on Flickr Database to testify our effectiveness in real-world application. The experiment results show that the Isomorphic Manifold Learning outperform state-of-the-art methods in keyvword annotation of images as well as the keyword-based image retrieval. Besides, it can deal well with the noise and redundancy in textual descriptors, and discover the latent associations between annotated keywords.

For future work, the incremental model for Isomorphic Manifold learning would be considered, and the training efficiency would be further improved.

7. ACKNOWLEDGEMENT
This research is supported by State 863 High Technology R&D Project of China (No. 2006AA01Z197), Program for China New Century Excellent Talents in University (NCET-05-03 34), Natural Science Foundation of China (No. 60775024).

8. REFERENCES