

An Adaptive Graph Model for Automatic Image Annotation*

Jing Liu
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, China
+86-10-62542971
jliu@nlpr.ia.ac.cn

Mingjing Li, Wei-Ying Ma
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, China
+86-10-58968888
{mjli, wyma}@microsoft.com

Qingshan Liu, Hanqing Lu
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, China
+86-10-62542971
{qslu, luhq}@nlpr.ia.ac.cn

ABSTRACT

Automatic keyword annotation is a promising solution to enable more effective image search by using keywords. In this paper, we propose a novel automatic image annotation method based on manifold ranking learning, in which the visual and textual information are well integrated. Due to complex and unbalanced data distribution and limited prior information in practice, we design two new schemes to make manifold ranking efficient for image annotation. Firstly, we design a new scheme named the Nearest Spanning Chain (NSC) to generate an adaptive similarity graph, which is robust across data distribution and easy to implement. Secondly, the word-to-word correlations obtained from WordNet and the pairwise co-occurrence are taken into consideration to expand the annotations and prune irrelevant annotations for each image. Experiments conducted on standard Corel dataset and web image dataset demonstrate the effectiveness and efficiency of the proposed method for image annotation.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval Models

General Terms

Algorithms, Experimentation,

Keywords

Image annotation, Image retrieval, Manifold ranking

1. INTRODUCTION

Nowadays, the digital images have become widely available on World Wide Web, which has brought about great challenges for efficiently searching and browsing the huge volume of available information. The key problem to design a successful image retrieval system is how to organize and rank the images according to the users' understanding on semantics, i.e. how to establish the connection or correspondence between image visual content and semantic description.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MIR'06, October 26-27, 2006, Santa Barbara, California, USA.
Copyright 2006 ACM 1-59593-495-2/06/0010...\$5.00.

The initial image retrieval is based on keyword query, that is, Query-By-Keyword (QBK), which is a simple extension of text retrieval. As the keyword annotation is done manually, this approach encounters the problem of inconsistency and subjectivity among different annotators, and the process is time-consuming as well. Especially with the explosive increase of images available, manually annotating all images is impractical. In early 1990's, an alternative scheme, content-based image retrieval (CBIR) was proposed. It takes example image as query (Query-By-Example, QBE) and rank images based on visual similarities. However, due to the well-known semantic gap, the performance of CBIR is far from satisfaction. Compared with QBE, the QBK method is more convenient and straightforward for users, in that the keyword can reflect some high-level semantics. Striving to overcome the aforementioned limitations, many researchers have devoted to realize automatic image annotation, which can be learned from some already annotated information or from textual information available in hosting web pages, such as image filenames, image caption and the surrounding text [11]. If automatic annotation can be achieved, the problem of image retrieval will be simplified into one of text retrieval, and many well-developed textual retrieval algorithms can be easily applied to find images by ranking the relevance between image annotations and textual queries. Thus, how to efficiently annotate the images is becoming one key issue for image retrieval, especially for QBK based retrieval.

In this paper, we focus on the task of automatic image annotation. We propose a novel method based on manifold ranking algorithm, in which the visual and textual information are well integrated. The original idea of manifold learning is to explore the intrinsic dimension and relations among data, and it is solved by constructing and analyzing a similarity graph of data. Due to its good properties of feature representation, it has been successfully applied in many applications, such as face recognition [8], graphics [23], document representation [9], and image retrieval [10]. Different from these works, we explore the manifold learning for image annotation. We make use of the good property of manifold learning that the label can be propagated through the similarity graph, so the annotations of unlabeled data can be learnt from its labeled neighbors. Moreover, the similarity graph can be regarded as a bridge connecting the semantic keyword space with the low-level feature space.

* This work was performed at Microsoft Research Asia.

Although the similarity graph is at the heart of graph-based method, its construction has not been studied extensively. A good graph can reflect a deep understanding of the data structure and help to learn valuable knowledge from the unlabeled data as much as possible. In the existing manifold studies, the similarity graph is often constructed by a simple k-NN or ϵ -ball method. Obviously, it is not suitable for the task of image annotation in large database, due to complex and unbalanced data distribution and limited prior information. Here, we design a novel method named the Nearest Spanning Chain (NSC) to construct the similarity graph, which is easy to implement. Furthermore, the method can leverage the imbalance and complexity in the real data distribution, and can ensure the less intra-similarity and larger inter-similarity. Owing to the huge amount and asymmetry characteristics of web images, this method is especially useful for web image annotation.

In the web environment, the initial annotations of images are either too few or too noisy. Thus the propagated keywords may be inaccurate. To solve this problem, we consider two kinds of semantic similarities among keywords. One is the correlation on semantic meaning and the other is the pairwise co-occurrence in images. [26] is the first attempt to improve the annotation accuracy by using the semantic similarity between keywords. It uses WordNet to capture keywords' similarity on semantic meaning. However, it doesn't consider word co-occurrence in each image. Generally, multiple keywords collectively describe different constituents in one image, for example, 'sky' and 'cloud', 'sea' and 'beach', 'polar' and 'bear'. Even if the co-occurring words are different in semantic meaning, they have close connection as the compound annotations for some images. Thus in order to annotate images properly and completely, we consider the co-occurrence relation for some keywords to establish the pairwise similarity measures, as well as utilize the result of WordNet.

In the following, we highlight the main contributions of our approach:

Firstly, we present a new image annotation method based on manifold ranking algorithm, in which visual and textual information are well integrated. Moreover, our graph-based method provides a very natural way to incorporate multi-modality information, such as the useful textual information for web images.

Secondly, we propose a novel method named NSC for the similarity graph construction. The NSC-based algorithm is easy to implement and robust across complicated data distribution.

Thirdly, we consider the correlation of keywords to further enhance the performance of annotation. Besides using the similarity relations obtained from the WordNet, we also employ pairwise co-occurrence of keywords. The pairwise similarity can be used to dynamically remove noisy keywords and expand the existing annotated keywords for image annotation.

The rest of this paper is organized as follows. Related works are briefly reviewed in Section 2. The framework of image annotation based on manifold ranking is presented in Section 3. We propose the method of NSC based adaptive similarity graph construction in Section 4. The construction and the usage of the pairwise word similarity are discussed in Section 5. The experimental results are reported in section 6. Finally, the conclusion and future work are given in Section 7.

2. RELATED WORK

Automatic image annotation is a key issue of QBK based image search. In recent years, a lot of algorithms have been presented for image annotation, which can be classified into three categories, i.e. classification method, probabilistic modeling method and graph-based method.

In the classification method, each annotated word is treated as an independent class and one semantic keyword corresponds to one classifier. The representative works are automatic linguistic index for pictures [15], content-based annotation method with SVM [2] and Bayes Point Machine [4], estimating the visual feature distributions associated with each keyword [6]. Because each semantic keyword usually should have one classifier, this type of method is unscalable for huge amount of images with infinite semantics.

Another direction is to learn the association probabilistic model between images and keywords. The early notable work is based on Translation Model (TM) [17] proposed by Duygulu et al., which applies a classical statistical machine translation model to translate the set of keywords for an image to the set of blob tokens obtained by clustering image regions. Another way of capturing co-occurrence information is to introduce latent variables to link image features with keywords. Three hierarchical probabilistic mixture models for image annotation [5] fall into this type, such as Gaussian Mixture Model, Latent Dirichlet Allocator (LDA) and correspondence LDA. Jeon et al [12] introduced a Cross-Media Relevance Model (CMRM), which used the keywords shared by the similar images to annotate new images. The CMRM method was subsequently improved through the continuous-space relevance model (CRM) [19] and the multiple Bernoulli relevance model (MBRM) [20]. All above statistical methods require tedious parameter estimation process and have the highly unbalanced components between visual features and text features, i.e. one image possesses continuous content feature vectors and very sparse observations in textual space.

Recently, graph-based method turns up to efficiently solve various machine learning problems, and it can also be imported as a meaningful application for image annotation. Pan et al [18] firstly proposed a Graph-based Automatic Caption (CGap) method. In their work, all images, as well as annotations and regions are represented as three types of nodes, and linked together according to their known association into one graph. This method has the advantages of being domain independent and simple parameter tuning, which are strong points shared by general graph model method. However region-based visual features are sampled from continuous sources and annotations are sampled from discrete sources of finite alphabet. It is difficult to weight these two types of nodes from different modalities in one graph. Furthermore, they do not consider the relationship among annotated keywords.

3. IMAGE ANNOTATION BASED ON MANIFOLD RANKING

In this paper, we propose a new image annotation method based on manifold ranking, in which the visual and textual information are well integrated. Before presenting the proposed framework of image annotation, we first introduce the manifold ranking algorithm for easily understanding the proposed method.

3.1 Manifold Ranking Algorithm

The manifold learning algorithm [24, 25] is initially proposed to explore the intrinsic dimension and property of the data. In this paper, we try to rank the data points or to predict the labels of unlabeled data points along their underlying manifold.

Assume a set of points $\chi = \{x_1, x_2, \dots, x_N\} \subset R^m$ (where m is the dimension for data features) and a label set $\zeta = \{1, 2, \dots, c\}$, in which part of data are labeled as $y \in \zeta$. Define a $N \times c$ matrix R corresponding to the ranking order for every label on the dataset, and a $N \times c$ initial labeling matrix $Y = [Y_1, Y_2, \dots, Y_c]$, with $Y_{ij} = 1$ if x_i is initially labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise. The manifold ranking procedure can be summarized as follows:

Step 1: Sort the pairwise distances among points in an ascending order. Repeat connecting the two points with an edge according to the order until a connected graph is obtained.

Step 2: Form the similarity matrix W defined by

$$W_{ij} = \exp\left[-d^2(x_i, x_j) / 2\sigma^2\right], \quad (1)$$

if there is an edge linking x_i and x_j . Let $W_{ii} = 0$.

Step 3: Symmetrically normalize W by

$$S = D^{-1} W D^{-1}, \quad (2)$$

where D is a diagonal matrix with diagonal element D_{ii} equal to the sum of the i^{th} row of W .

Step 4: Do iteration according to Equation 3, until convergence.

$$R(t+1) = \alpha \times S \times R(t) + (1-\alpha) \times Y, \quad (3)$$

where t represent the number of iterations, $\alpha \in [0, 1]$ is the propagation parameter and $R(0) = Y$.

Step 5: Output the ranking result R^* , whose element is the ranking value corresponding to each label for each point, i.e. R_{ij}^* is the ranking value to label j for point x_i .

3.2 The Framework for Image Annotation

The proposed annotation scheme based on manifold ranking includes two main components, i.e., adaptive similarity graph construction and dynamic adjustment of the labelling matrix. The framework of the proposed method is shown in Figure 1.

In terms of the construction of the graph model, we explore the visual content-based similarity relationship among all the images in the dataset and propose a novel method to weight the similarities. Then we obtain one updated similarity matrix S^* . The detailed algorithm is described in Section 4.

In addition, the semantic similarities between any two keywords are obtained by the WordNet and statistical co-occurrence information. Here, the word-to-word correlation has two usages. First, we multiply the labeling matrix Y in Equation 3 by the correlation matrix K^* . Accordingly more related keywords can be linked together through expanding the semantic meaning for each annotated image. Second, we measure pair-wise similarities for the predicted annotations of each un-annotated image and prune some irrelevant keywords. The details will be presented in Section 5.

After we get the weighted similarity graph model S^* and updated labelling matrix Y^* , we can perform the iterative procedure according to Equation 3. Finally, we can get different ranking values corresponding to every keyword for each image. Here, we select top- n keywords as the candidates (n is more than common required annotation length), then prune some irrelevant ones by using the word-to-word correlation to obtain the final annotations for each unknown image.

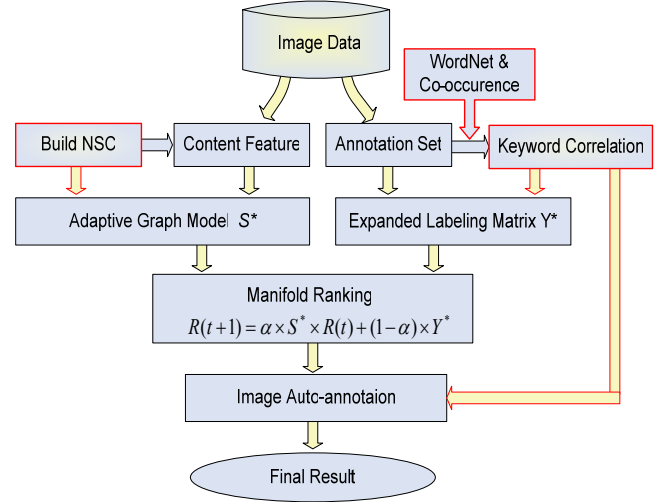


Figure 1. Processing Flow of Auto-Annotation

4. ADAPTIVE GRAPH MODEL

In the previous manifold studies, most of them just use simple methods for graph construction, such as full-connection graph, ϵ -ball or k -NN methods. Image space might be a non-linear sub-manifold, which is embedded in the ambient space [22]. Moreover, for image annotation, there is a large amount of data needing to be represented in the graph. Thus, ϵ -ball or k -NN based method is more suitable than the full-connection method. However, these popular methods depend on parameter ϵ or k . Unsuitable parameter tuning would result in shortcuts or bad connections among data. In order to overcome the problem existing in ϵ -ball or k -NN based method, we propose the NSC based method for graph construction.

4.1 Nearest Spanning Chain

The Nearest Spanning Chain (NSC) method is inspired by a graphic concept denoted as the Minimum Spanning Tree (MST). Similarly, both of them connect all the points into a graph. However, it is different from the MST. The NSC is a sequential chain, which has multiple one-to-one connections without any branches, while the MST tree usually has some branches and one-to-many connections. So the proposed NSC is much simpler than the MST. In the following, we define the properties of the Nearest Spanning Chain and give a simple example of NSC shown in Figure 2:

Property 1: N points (2 end points and $N-2$ mid points) and $N-1$ edges form one sequential chain.

Property 2: Two end points are connected with one edge respectively and other mid points are connected with two edges.

Property 3: The points in the chain are connected by the *Nearest in Residue* (NinR) rule.

where the rule of NinR means that every point should find its nearest neighbor in residual un-selected points.

Although above three properties cannot ensure every edge in NSC being the nearest connection, two nearer (more similar) points should have high probability to be connected together and two farther (more un-similar) points have low probability to be connected. Therefore, if we use multiple NSCs, their statistical information can reflect the data distribution to some extent. We can randomly select one point as the starting point each time to build one NSC, and repeat this process to build multiple NSCs. In addition, considering that the NSC is a sequential chain, the connection occurring in front part should have more reliability to be one nearer pairwise connection. Therefore, according to the occurring sequence for every connection, we use the weighted sum of all the pairwise connections as the statistical measure. Generally, the weighting coefficients should be inversely proportional to the occurring position of every pairwise connection in the sequence.

4.2 Graph Model Construction

We try to construct the similarity graph model for manifold ranking with better robustness and lower computation cost, when the distribution of data points is very complex or even unbalance. Based on the statistical measures from multiple NSCs, we can achieve an adaptive construction for the graph model. The algorithm can be simply understood as weighting the original similarity matrix according to the data distribution.

Assume that there have N points in the dataset and n NSCs are found (usually $n \ll N$, proved in the following experiment). The procedure of the adaptive graph model construction is as follows:

Step 1: Randomly select one point in the whole dataset and take it as the starting point to build one NSC.

Step 2: The nearest neighbor to the prior selected one is selected as the next starting point for the next nearest neighbor search.

Step 3: Repeat the nearest neighbor's search in unselected residual points, until all the points in the dataset have been selected

Step 4: Turn to the first step and building another new NSC until we have obtained n sequential chains.

Step 5: Count the same pairwise connections in obtained n chains and build a connection statistical matrix C ($N \times N$), where C_{ij} is the weighted sum of connection occurrence for i^{th} point and j^{th} point ($C_{ij} = C_{ji}$).

$$C_{ij} = \gamma \sum_{m=1}^n seq_w_{ij}^m \cdot \delta_{ij}, \quad (4)$$

where γ is a parameter in $(0,1)$ to adjust the NSC's impact on the construction of similarity graph. Empirically, we set $\gamma = 1/n$. $\delta_{ij} = 1$ when i^{th} point is connected with j^{th} point, otherwise $\delta_{ij} = 0$. $seq_w_{ij}^m$ is the sequential weight for the

connection between i^{th} point and j^{th} point occurring in m^{th} chain. In the experiments, due to the huge amount of points, we would cut the long chain into short parts with equal length, and use the reverse indexes of obtained parts to weight every pair-wise connection.

Step 6: Construct the similarity matrix W as Equation 1.

Step 7: Obtain a new similarity matrix W^* by using C to weight W as follows:

$$W_{ij}^* = c_{ij} \exp\left[-d^2(x_i, x_j) / 2\sigma^2\right], \quad (5)$$

Step 8: Compute S^* with W^* according to Equation 2 to prepare for the manifold ranking.

4.3 Toy Example

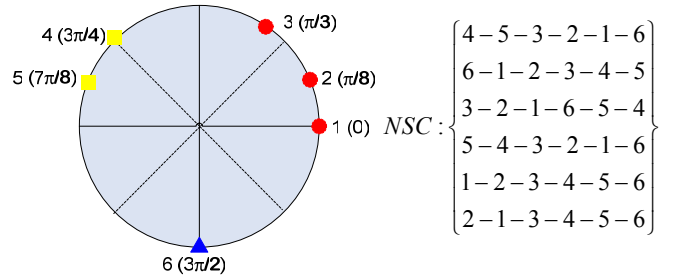


Figure 2. Toy Dataset & NSCs: the left in figure presents the data distribution. The right in figure gives six examples of NSC denoted by the indexes of data.

Here, we take one toy example shown in Figure 2 to explain the principal of this adaptive similarity graph construction. In Figure 2, the indexes and coordinates for each point are denoted by the numbers outside of bracket and in the bracket respectively. The toy data are along a circle, i.e., they distribute on a simple manifold. Given three data sets belonging to three different classes, which are denoted as A $\{0, \pi/8, \pi/3\}$, B $\{3\pi/4, 7\pi/8\}$, C $\{3\pi/2\}$ respectively, we build the NSC according to description in the subsection 4.2. The data sets and obtained multiple NSCs are shown as Figure 2, where the distance measure is the shortest distance along the circle.

Firstly we generate random numbers from 1 to 6 to decide the sequential indexes of starting points, such as (4 6 3 5 1 2). Next, we select the point $4(3\pi/4)$ in B as the starting point, and find its nearest neighbor, the point $5(7\pi/8)$. Excluding the searched point 4, the nearest neighbor of point 5 is point $3(\pi/3)$, and $2(\pi/8)$, $1(0)$, $6(3\pi/2)$ in turn. In this way, we get one NSC (4-5-3-2-1-6). In the following, we select another starting point $6(3\pi/2)$ to get another NSC (6-1-2-3-4-5). Then repeat the same process for other starting points. Here we make every data as starting point once, since the number of data is very small. In practice, when we have thousands or more than thousands of data points, the amount of NSCs can be much less than the number of data, which is proved through the latter experiment in section 6. Finally we get the weighted occurrence for every pairwise connection and combine the occurrence matrix with the similarity matrix (Here, we simplify the similarity matrix with the reciprocal items of distance matrix and normalize the minimum distance $\pi/8$ to be 1).

According to above designment, similarity matrix (W) and weighted similarity matrix (W^*) are given as follows ($W_{ii} = 0$):

$$W = \begin{bmatrix} 0 & 1 & \frac{3}{8} & \frac{1}{6} & \frac{1}{7} & \frac{1}{4} \\ 1 & 0 & \frac{3}{5} & \frac{1}{10} & \frac{1}{13} & \frac{1}{20} \\ \frac{3}{8} & \frac{3}{5} & 0 & \frac{3}{10} & \frac{3}{13} & \frac{3}{20} \\ \frac{1}{6} & \frac{1}{10} & \frac{3}{10} & 0 & 1 & \frac{1}{6} \\ \frac{1}{7} & \frac{1}{13} & \frac{3}{13} & 1 & 0 & \frac{1}{5} \\ \frac{1}{4} & \frac{1}{20} & \frac{3}{20} & \frac{1}{6} & \frac{1}{5} & 0 \end{bmatrix} \quad W^* = \begin{bmatrix} 0 & \frac{11}{3} & \frac{1}{4} & 0 & 0 & \frac{5}{12} \\ \frac{11}{3} & 0 & \frac{3}{2} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{2} & 0 & \frac{3}{5} & \frac{2}{13} & 0 \\ 0 & 0 & \frac{3}{5} & 0 & \frac{8}{3} & 0 \\ 0 & 0 & \frac{2}{13} & \frac{8}{3} & 0 & \frac{2}{15} \\ \frac{5}{12} & 0 & 0 & 0 & \frac{2}{15} & 0 \end{bmatrix}$$

Note that in this example the weights for pairwise connection occurrence matrix are simply set from 5 to 1 in the ascending order of the indexes for each connection, i.e. the connection between the 1st and 2nd point is weighted by 5, and the connection between the 5th and 6th point is weighted by 1, and the rest may be deduced by analogy. In the real experiments, we should weight each pairwise connection according to the description in subsection 4.2.

From the above toy example, it can be found that the connection inside one dataset become denser (such as point 1 and point 2 or 3) and the connection between different datasets become sparser (such as point 1 and point 4 or 5). That is to say, the proposed method can adapt to the data distribution to some extent.

5. THE CORRELATION OF KEYWORDS

The other main component of the proposed framework shown in Figure 1 is the utilization of word-to-word correlations. It is used in two ways. One is to adjust the labeling matrix Y , and the other is to refine predicted annotations. The correlation is got by analyzing the relations of keywords. In this paper, we consider two kinds of correlation information, i.e., correlation by the WordNet and correlation by statistical co-occurrence.

5.1 Correlation by WordNet

In WordNet [16], synonyms with the same meaning are grouped together to form a synset called as concept. Various concepts are linked to each other through different relations including **hyponyms**, **part of** and **member of**. Therefore, WordNet is useful to determine semantic connections between sets of synonyms. We will use the structure and content of WordNet for measuring the pairwise semantic similarity.

The measures based on WordNet can be classified into three different categories: node-based, distance-based and gloss-based. The Jiang and Conrath Measure (JNC) [13], which integrates the node-based and edge-based approach together, is proved to be most effective one to measure the semantic distance between two concepts [1]. Given two words w_i and w_j , we should firstly find its associating concepts c_i and c_j (Note that a keyword may be associated with more than one concepts in WordNet), and get the maximum similarity among all possible corresponding concept-pairs as the semantic similarity for the two words.

$$K_N(w_i, w_j) = \max_{c_i, c_j} Sim(c_i, c_j) \quad (6)$$

$$Sim(c_i, c_j) = \frac{1}{IC(c_i) + IC(c_j) - 2 * IC(lcs(c_i, c_j))} \quad (7)$$

$$IC(c_i) = -\log \left[\frac{freq(c)}{N} \right] \quad (8)$$

where $lcs(c_i, c_j)$ is the lowest common subsumer between concept c_i and c_j , $freq(c)$ is the appearance frequency for all the words in the Corpus associated to the concept c , and N is a normalizer.

5.2 Correlation by Co-occurrence

Strictly speaking, similarity measure of WordNet only considers the hyponymy hierarchy. It does not measure more general semantic relatedness. Statistical co-occurrence for annotated keywords in images is an important part of semantic relatedness for image annotation. Intuitively, two keywords with high co-occurrences will lead to high probability for annotating one linked by another.

In this paper, we count the frequency of every keyword-pair simultaneously to be annotations for one image and obtain the keyword co-occurrence matrix $K_C (M \times M)$, where M is the total number of keywords. $K_C(w_i, w_j)$ is the frequency of co-occurrence for keyword w_i and w_j . Generally speaking, the more common a keyword is, the more chance it will associate other keywords. However this kind of associations has lower reliability. Therefore, we weight the counts according to the uniqueness of each keyword, i.e. setting a lower weight to frequent keywords and a higher weight to unique keywords. The weighted keywords co-occurrence matrix K_{Cw} can be calculated as follows:

$$K_{Cw}(w_i, w_j) = K_C(w_i, w_j) \times \log \left(\frac{N}{n_i} \right) \quad (9)$$

where n_j is the frequency for keyword j occurring in the annotated images, and N is the total number of images in the dataset.

Note that $K_{Cw}(w_i, w_j)$ is usually not equal to $K_{Cw}(w_j, w_i)$, i.e. when w_i occurs, w_j has high probability to occur, the converse case is not true.

5.3 Combined Correlation

After obtaining above two types of semantic similarity, we firstly normalize them into $[0, 1]$ and then combine them linearly.

$$K^*(w_i, w_j) = \eta K_N(w_i, w_j) + (1 - \eta) K_{Cw}(w_i, w_j) \quad (10)$$

where η and $(1 - \eta)$ are weights for the combination to achieve the complement each other. In our experiment, we set η to 0.5 as default.

With the obtained correlation matrix K^* , we would utilize it into our framework for image annotation. Before the propagation through manifold ranking, K^* is used to expand the existing annotation for each annotated image by the following equation.

$$Y^* = Y \times K^*, \quad (11)$$

where Y is the labeling matrix mentioned in Equation 3.

After finishing the manifold ranking, top possible keywords can be obtained from the matrix R in Equation 3. However, these keywords cannot ensure the consistence on semantics, such as ‘sea’ and ‘beach’ are more possible to appear simultaneously than ‘sea’ and ‘street’. Therefore, we can further utilize the word

correlation matrix K^* to remove irrelevant keywords from the returned ones. Here the removal is decided by the sum of mutual similarity measures for one keyword with other returned keywords.

$$Sum_i = \sum_{j \in R^i} K_{ij}, \quad (12)$$

where R^i is denoted as the set of returned keywords except keyword i . Then the keywords whose sum value are larger than a certain threshold or top m keywords become the final annotations.

6. EXPERIMENTAL RESULTS

6.1 Experiment Design

We test the proposed algorithm on three datasets: two Corel data sets from Duygulu et al. [14, 17] and one web image dataset from Tong et al. [22].

Corel Dataset (ECCV): The dataset is extensively used as basic comparative data for recent research work in image annotation. There are 5000 images from 50 Stock Photo CDs in this dataset. Each CD contains 100 images on the same topic. We use the normalized cut algorithm [21] to segment 5,000 images, and totally obtain 47,065 segmented regions. For each image, we select at most 10 largest regions and extract 30-dimensional low-level features from each region. Each image is annotated with 1 to 5 words and totally 371 words have been used in annotations. For the convenience of comparison, 4,500 images out of 5,000 are previously labeled and the rest 500 hold-out images are used for testing, in which totally 260 distinct words are found.

Corel Dataset (JMLR): These are 10 image data sets from Corel. On average, each data set has around 5,000 images totally including 50,000 regions and 165 words in the caption vocabulary. The structure and extracted features are the same as that of Corel dataset (ECCV).

Web Dataset: 9,046 web images are crawled from web. Every image is annotated by 5 to 10 keywords from the surrounding text and tag information, which is extracted from the blocks containing the images by the VIPS algorithm [3] and processed by the standard text processing technique. There are totally 1,153 keywords excluding some rare ones, i.e. occurring frequency is less than 5. Additionally, 144 dim of low-level features are extracted from each image, including 36 dim color histogram, 64 dim color correlogram, 20 dim Tamura feature and 24 dim pyramid wavelet texture feature.

Because the annotations for the Corel dataset are given manually, they have relatively good correspondence to regions. Accordingly, when it is used, we build the region-based similarity graph model. Then the ranking value for each keyword corresponds to each region. Here we select top 3 annotations per region in an image as candidates for this image’s annotations, and decide the final top 5 annotations according to the pairwise word correlations discussed in Section 5.2. In addition, the initial labels for every region are defined as the labels for corresponding image.

For web dataset, due to the limitation on annotation extraction for web images, the obtained annotations are usually diverse and inconsistent with region’s semantics. Then we build the similarity graph model based on the global content features for web images. The ranking value for each annotation is directly given to an image. Here we select top 10 annotations per image and decide the final top 5 annotations with consideration to word correlations.

Similar to previous works on image annotation, the quality of automatic image annotation is measured by the performance of retrieving auto-annotated test images regarding to single-word queries. For each single-word query, the number of correctly annotated images is denoted as N_c , the number of searched images is denoted as N_s , and the number of related images in dataset is denoted as N_r . Then the precision and recall are computed as follows:

$$precision(w) = \frac{N_c}{N_s}, \quad recall(w) = \frac{N_c}{N_r} \quad (13)$$

We compute the average precision and recall over all words occurring in test images to evaluate the performance of the image auto-annotation method.

Another evaluating measure is the annotation accuracy, which is the percentage of correctly annotated keyword (M_c) in all returned ones (M_r) for an image. Additionally, the length of returned annotations is the same to ground truth for the image. The average accuracy (A_c) over all test images is the final measure.

$$A_c = \frac{M_c}{M_r} \quad (14)$$

6.2 Experiment I: Adaptive Graph Model vs. k -NN based Graph Model

In experiment I, we use Corel dataset (ECCV) to compare the proposed NSC-based method (noted as N-Precision & N-Recall) with the k -NN based method (noted as K-Precision & K-Recall). Here we also combine the word correlation into the procedure of image annotation.

We set k to 10, 20, 30, 40 and 50 respectively to construct k -NN similarity graph. For the proposed method, we use 50-NN similarity matrix, and 100, 200, 300, 400 and 500 NSCs respectively to build the adaptive graph model. α in Equation 3 is set to 0.15 (The following experiments also have the default value). The average precision and recall for both methods over all the annotated 260 words are respectively shown in Figure 3 and Figure 4.

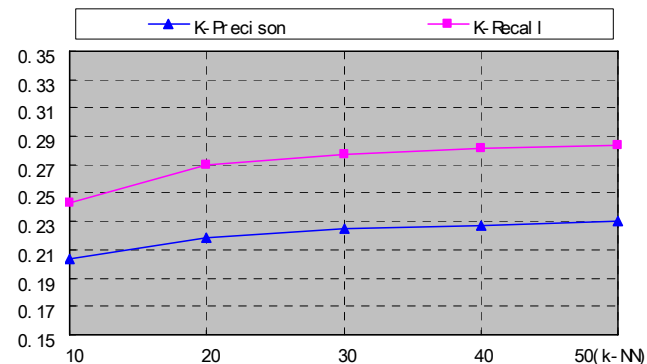


Figure 3. Average Precision & Recall for different k values in k -NN based method.

Observed from curves in Figure 3, small deviations around $k=40$ make little difference. Accordingly, we use the measures at 40-NN as the baselines, which are shown as dashed line in Figure 4. From Figure 4, we can see that the NSC based model shows better performance than k -NN based model in both the average precision

and recall, especially in precision. Because the NSC based model adaptively weights the pair-wise similarities, considering the real distribution for high dimensional content features rather than only using the same number of neighbours for each node in the graph, it can achieve the better performance on the semantic propagation. With the increasing number of built NSCs (n -NSC), the performance is also enhanced. While the number is around 400, the trend becomes smooth gradually. Thus in the following experiments on Corel datasets, we set 400 as the default value to construct the NSC based graph model, and k is 50.

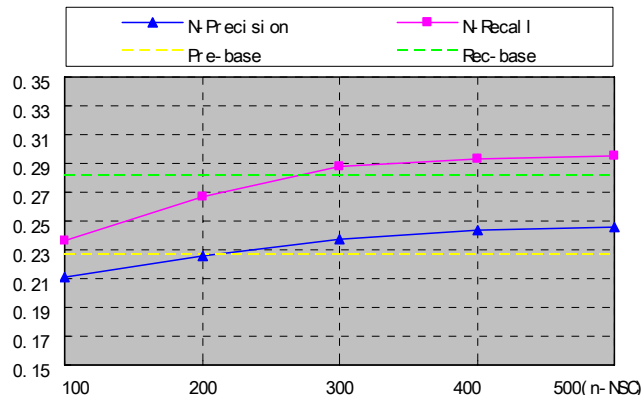


Figure 4. Average Precision & Recall for different number of NSCs in our method: the solid lines are results from our method, and the dashed lines are corresponding baselines obtained from k -NN based method.

6.3 Experiment II: Comparison with Other Related Works

In the subsection, we compare the proposed method with other related works, in order to further show the advantage of the proposed method.

Firstly, we compare our adaptive graph-based annotation method (noted as AGAnn) with the GCap (Graph-based Automatic Captioning) method [18]. Same as [18], we test two methods on the Corel dataset (JMLR), and evaluate the methods with the criterion in Equation 14. Figure 5 reports the results on all the ten data sets, where the results of the GCap are from [18].

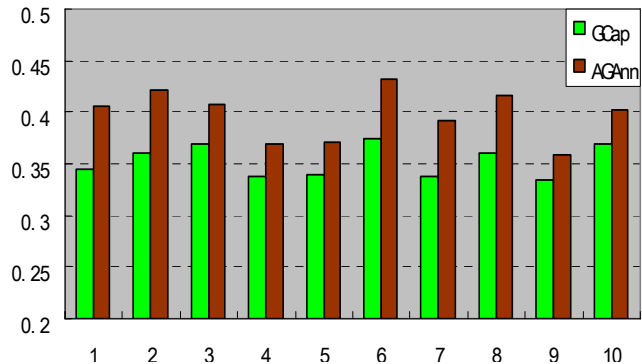


Figure 5. Annotation Accuracy Comparison in Experiment II

As shown in Figure 5, our proposed method achieved improvement on all of the ten data sets. The improvement can be analysed from two aspects. One is that the similarity cannot be effectively propagated in the GCap, for it puts two feature elements with different modalities directly into one graph. While

our method makes the semantic information direct the propagation and considers the data distribution more suitably. Another is that the import of word correlations can prune some irrelevant returned annotations so as to enhance the annotation accuracy.

Another related work is Yohan Jin’s TMHD [26]. As we discussed in Section 1, TMHD also considered the pairwise word relation and integrated it into the removal of irrelevant annotation keywords. We employ Corel dataset (ECCV) and evaluation measures, which are same to Yohan Jin’s. Here we stress on comparing two different manners to get and use the word correlations. Furthermore, the improvement brought by word correlation cannot evidently exert on infrequent keywords. Therefore, we only report experimental results based on seven most frequent keywords, which are same as TMHD’s. Additionally, we use the TM’s results as a baseline. The results for TMHD and TM are directly referred to [26].

As Table 1 shows, compared with TM, our method (AGAnn) gains improvement both on recall and on precision, while TMHD gains increased precision but losses on recall. This happens due to the wrong removal of keywords in returned annotations, i.e. the word correlation obtained only by WordNet cannot provide sufficient guide. Moreover, in the region-based annotation method, the semantic relation among segmented regions can be shown more reasonable on the word co-occurrence than on the keyword meaning.

Table 1. Performance of Most Frequent Keywords for TM, TMHD and AGAnn

Keyword	TM		TMHD		AGAnn	
	Prec.	Recall	Prec.	Recall	Prec.	Recall
nest	0.1250	0.1428	1.0000	0.1428	1.0000	0.1428
tiger	0.1428	0.3000	0.5000	0.1000	1.0000	0.1000
stone	0.1666	0.3809	0.1702	0.3809	0.5000	0.3809
water	0.2482	0.8965	0.5000	0.0413	0.2410	0.9741
plane	0.1428	0.1600	0.1481	0.1600	0.5284	0.3600
window	0.1111	0.1250	0.1111	0.1256	0.2000	0.2000
garden	0.0952	0.2000	0.1666	0.1000	0.1250	0.1000
Aver.	0.1474	0.3150	0.4049	0.1501	0.5135	0.3225

6.4 Experiment III: Evaluation on Web Image Dataset

Generally, web images have extensive semantics and large variation on visual content. Then the auto-annotation for web images becomes a great challenge and highly meaningful work. Provided with the property of similarity propagation, the graph-based method usually requires less labelled data and can achieve comparatively satisfactory accuracy, which we can see from Figure 6 for Corel dataset (ECCV). Thus we extend our work to web image annotation and test its performance on web images as mentioned in Section 6.1.

In this experiment, we randomly select images from the dataset by different percentages as the labeled ones and the rest of images form the test dataset. The evaluation measures used in this paper are according to Equation 13. Figure 7 shows the experimental result by AGAnn. Then can see that the performance of AGAnn is also insensitive to the labeled image percentage, which make the













					
Butterfly Animal Nature Flower Sun	Blue Whale Life Mammal Sea	Plumage Ptarmigan Male Bird Nature	Mountain Pine Woodland Park Forest	Flower Yucca White Nature Tree	Man Soil Nature Mountain Sky
					
Tom Leaf Boyden Bumblebee Flower	Dolphin Marine Mammal Croatia Water	Earth Science Planet Water Ball	Space Broadcast Satellite Circle Search	Nature Crane Enchant Dance Mate	Oriol Hood Stone Wood Male

Figure 8. Examples for Web Image Annotation

annotation for huge volume of images possible. In Figure 8, we give some examples of image annotation result, in which the annotations for each image are basically consistent in the semantic relevance, even though they sometimes don't properly annotate the image.

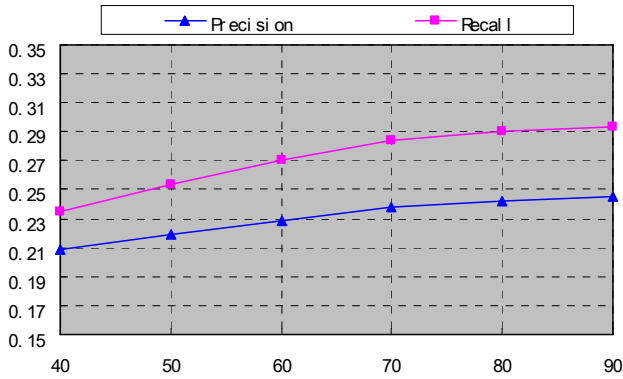


Figure 6. Average Precision & Recall for Corel dataset (ECCV): The horizontal axis denotes the percentage of annotated images. In all cases, AGAnn uses $\alpha = 0.15$, $k = 50$, $n\text{-NSC} = 400$.

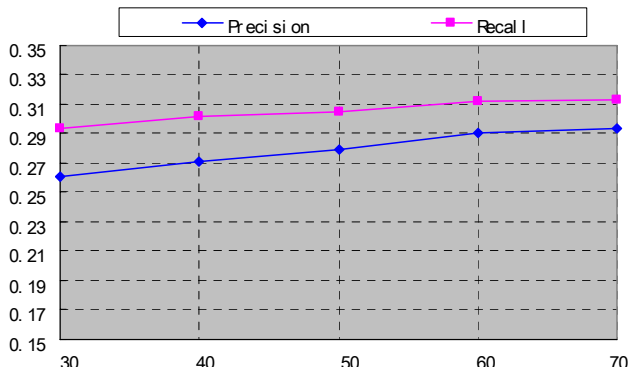


Figure 7. Average Precision & Recall for Web dataset: The horizontal axis denotes the percentage of annotated images. In all cases, AGAnn uses $\alpha = 0.1$, $k = 100$, $n\text{-NSC} = 500$.

Furthermore, it is worth mentioning that the performance of our system can be further improved. This is because the experiment is completely based on the original annotations in the web dataset, which are automatically extracted from corresponding hosting webpage without any human intervention. So some annotation as the ground truth in evaluation may be noise. This is just what we would try to improve in the future work.

7. CONCLUSION AND FUTURE WORK

In this paper, we develop an efficient framework based on manifold ranking algorithm to realize automatic image annotation. Our learning model is constructed in a straightforward manner by exploring the relationship among all images in the feature space and among all annotated keywords. The Nearest Spanning Chain (NCS) method is proposed to construct the similarity graph that can locally adapt to the complicated data distribution. Furthermore, by the usage of WordNet and statistical co-occurrence measure, the word-to-word correlations are successfully utilized to expand the semantic meaning for each annotated image and prune irrelevant annotations for each un-annotated image. Experiments on Corel datasets and web image dataset show encouraging performance of the proposed method.

Our goal is to automatically annotate a huge amount of images precisely and efficiently, and simultaneously require labeled information as less as possible. Hence in the future work, we will work on web image annotation by mining and refining more relevant semantic information from web pages and building more suitable connection between image content features and available semantic information.

8. REFERENCES

- [1] Budanitsky, A. and Hirst, G. *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*. In Workshop on WordNet and Other Lexical Resources, 2nd of the North American Chapter of the ACL, Pittsburgh, 2001.
- [2] Claudio, C., Gianluigi, C., Raimondo, S. *Image annotation using SVM*. In Proceeding Of Internet imaging IV, Vol. SPIE, 2004.

- [3] Cai, D., Yu, S., Wen, J.R. and Ma, W.Y. *VIPS: a vision-based page segmentation algorithm*. Microsoft Technical Report (MSR-TR-2003-79), 2003.
- [4] Edward Chang, Kingshy Goh, Gerard Sychay, Gang Wu. *CBSA: content-base soft annotation for multimodal image retrieval using bayes point machines*. *CirSysVideo*, pp. 26-38, 13(1), 2003.
- [5] David, M., Michael, I. Jordan. *Modeling annotated data*. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 127-134, July 2003.
- [6] Gustavo, C. and Nuno, V. *A database centric view of semantic image annotation and retrieval*. In ACM SIGIR Conf. on Information Retrieval, Salvador, Brazil. 2005.
- [7] He, J., Li, M., Zhang, H.J. Tong, H. and Zhang, C. *Manifold-ranking based Image Retrieval*. Proceedings of the 12th annual ACM international conference on Multimedia, pp. 9-16, New York, 2004.
- [8] He, X., Cai, D., Liu, H. and Ma, W.Y. *Learning a locality preserving subspace for visual recognition*. In Proc. IEEE Conf. on Computer Vision, Nice, France, 2003.
- [9] He, X., Cai, D., Liu, H. and Ma, W.Y. *Locality preserving indexing for document representation*. In ACM SIGIR Conf. on Information Retrieval, Sheffield, 2004.
- [10] He, X., Ma, W.Y. and Zhang, H.J. *Learning an image manifold for retrieval*. In Proc. of ACM international conference on Multimedia, New York, USA, 2004.
- [11] Hua, Z.G., Wang, X.J., Liu, Q.S. and Lu, H.Q. *Semantic knowledge Extraction and Annotation for web images*. Proceedings of the 13th Annual ACM International Conference On Multimedia, pp. 467-470, Singapore, 2005.
- [12] J. Jeon, V. Lavrenko and R. Manmatha. *Automatic Image Annotation and Retrieval Using Cross-media Relevance Models*. In Proc. of ACM SIGIR conference on Research and development in information retrieval, pp. 119-126, July 2003.
- [13] Jiang, J. and Conrath, D. *Semantic similarity based on corpus statistics and lexical taxonomy*. In Proceedings on International Conference on Research in Computational Linguistics, 1997.
- [14] Kobus, B., Pinar, D. et al. *Matching words and pictures*. *Journal of Machine Learning Research*, pp. 1107-1135, 2003.
- [15] Li, J. and Wang, J. Z. *Automatic linguistic indexing of pictures by a statistical modeling approach*. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, pp. 1075-1088, 25(19), 2003.
- [16] Pucher, M. *Performance Evaluation of WordNet-based Semantic Relatedness Measures for Word Prediction in Conversational Speech*. In Sixth International Workshop on Computational Semantics, Tilburg, Netherlands, 2005.
- [17] Pinar, D. and Kobus, B. *Object recognition as machine translation: learning a lexicon for a fixed image vocabulary*. In Seventh European Conference on Computer Vision, 4:97-112, 2002.
- [18] Pan, J.Y., Yang, H.J. and Pinar, D. *Automatic multimedia cross-modal correlation discovery*. The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 653-658, August 2004.
- [19] R. Manmatha, V. Lavrenko, and J. Jeon, *A Model for Learning the Semantics of Pictures*. In Proc. of the 17th Annual Conf. on Neural Information Processing Systems, 2003.
- [20] S. L. Feng, R. Manmatha and V. Lavrenko. *Multiple Bernouli Relevance Models for Image and Video Annotation*. In Proc. Of CVPR, Washington, DC, June, 2004.
- [21] Shi, J.B. and J. Malik. *Normalized Cuts and Image Segmentation*. *IEEE Conference Computer Vision and Pattern Recognition(CVPR)*, pp. 731-737, June 1997.
- [22] Tong, H., He, J., Li, M.J., Zhang, C., and Ma W.Y., *Graph Based Multi-Modality Learning*. Proceedings of the 13th Annual ACM international conference on Multimedia, pp. 862-871, Singapore, 2005.
- [23] Wojciech, M., Hanspeter, P., Matt, B. *A data-driven reflectance model*. In Proc. of SIGGRAPH, 2003.
- [24] Zhou, D., Bousquet, O., Lal, T.N., Weston, J., and Schölkopf, B. *Ranking on Data Manifolds*. 18th Annual Conf. on Neural Information Processing System, pp. 169-176, 2003.
- [25] Zhou, D., J. Huang and B. Schölkopf. *Learning with local and global consistency*. 18th Annual Conference on Neural Information Processing Systems, 2003.
- [26] Yohan Jin, Khan, L., Wang, L. and Awad, M. *Image Annotations By Combining Multiple Evidence & WordNet*. Proceedings of the 13th Annual ACM International Conference On Multimedia, pp. 706-715, Singapore, 2005.