Automatic Image Annotation via Local Multi-Label Classification

Mei Wang[1] Xiangdong Zhou[1,2] Tat-Seng Chua[2]
[1] Fudan University, Shanghai, China
[2] National University of Singapore, Singapore
{xdzhou, 051021052}@fudan.edu.cn, chuats@comp.nus.edu.sg

ABSTRACT

As the consequence of semantic gap, *visual similarity does not guarantee semantic similarity*, which in general is conflicting with the inherent assumption of many generative-based image annotation methods. While discriminative learning approach had often been used to classify images into different semantic classes, its efficiency is often impaired by the problems of multi-labeling and large scale concept space typically encountered in practical image annotation tasks. In this paper, we explore solutions to the problems of large scale concept space learning and mismatch between semantic and visual space. To tackle the first problem, we explore the use of higher level semantic space with lower dimension by clustering correlated keywords into topics in the local neighborhood. The topics are used as lexis for assigning multiple labels for unlabeled images. To tackle the problem of semantic gap, we aim to reduce the bias between visual and semantic spaces by finding optimal margins in both spaces. In particle, we propose an iterative solution by alternately maximizing the sum of the margins to reduce the gap between visual similarity and semantic similarity. The experimental results on the ECCV2002 benchmark show that our method outperforms the state-of-the-art generative-based annotation method MBRM and discriminative-based ASVM-MIL by 9% and 11% in terms of $F1$ measure respectively.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis; H.3.3 [**Information Storage and Retrieval**]

General Terms

Algorithms, Experimentation, Performance

Keywords

Automatic image annotation, multi-label classification, maximum margin clustering

1. INTRODUCTION

The growing number of digital images has brought about an urgent need to facilitate the retrieval and browsing of images via semantic keywords. Thus, techniques for Automatic Image Annotation (AIA) becomes increasingly important and a large number of machine learning techniques have been applied along with a great deal of research efforts. However, AIA task presents unique problems of multi-label classification [16] and large scale concept space [24]. These problems make AIA different and challenge for many traditional machine learning techniques and exacerbate the problem of *semantic gap* .

Many image annotation methods based on various learning techniques have been proposed [34, 13, 17, 9, 16, 15, 19, 8, 30, 1, 23, 10, 18, 20, 12, 21, 11, 2, 31, 27, 26]. We can roughly classify these methods into two categories. The first category treats each semantic keyword or concept as an independent class and trains a corresponding classifier based on the training set to identify images belonging to this class. The earlier efforts in this category were applied to extracting specific semantics, such as differentiating indoor from outdoor scenes [30], cities from landscapes [1], and detecting trees [23], horses [10], or buildings [18], etc.

The representative technique for the first category is the classification technique such as the Support Vector Machine (SVM), which demonstrates strong discrimination power. The problem with the classification-based techniques is that they are not very scalable to large scale concept space. In the filed of AIA, the semantic space is growing larger and larger along with more structural information. For instance, the widely used Corel data set contains more than 300 semantic labels [7], while the goal of LSCOM project is to build a semantic space consisting of thousands of concepts with rich semantic connections [24]. Therefore, the problem of semantic overlap and data imbalance among different semantic classes induced by the multi-label characteristics of AIA is becoming more serious. Consequently, the classification power of this kind of approach is heavily impaired. Other methods in this category include [34], [11].

The second category of AIA methods focuses on learning the correlations between the visual features and semantic concepts. Many such methods are based on the *generative model*, in which an influential work is CMRM [13], which tries to estimate the joint probability of the image's visual keywords and the semantic keywords on the training set. It was subsequently improved through a continuous relevance model [17], multiple Bernoulli relevance model [9] and the recently proposed dual cross-Media relevance model

[19]. There are also efforts to consider the keyword correlations in the annotation process, such as the Coherent Language Model (CLM) [14], the Correlated Label Propagation (CLP) [16], the WordNet-based method [15, 27] and the keyword correlations based concept annotation [36]. The Web sources are also exploited to improve image annotation [19]. Recently, Qi et al. [25] proposed a correlative multi-label (CML) annotation framework which simultaneously classifies concepts and models their correlations for video annotation.

The generative-based (visual features & keywords) methods have shown better durability to the scalability of concept space, and provides a natural ranking for choosing the proper keywords as semantic annotations. However, many such methods are based on the strong assumption that *visual similarity guarantees semantic similarity* which is often violated as a consequence of the well-known *semantic gap* problem. For instance, images belonging to the same visual neighborhood often do not share similar semantic contents. In fact, the *semantic gap* problem implies that similar visual contents may correspond to multiple different semantic meanings. It is one of the reasons that the intuitive approach of designing a "good" metric measurement or density estimation method to directly bridge the semantic gap does not lead to satisfactory results.

The above discussions highlighted the key challenges improving the performance of AIA task. The challenges includes (1) the ability to scale up the large concept space, and (2) the mismatch between visual similarity and semantic similarity. To handle the first problem, we explore higher level (lower dimensional) semantic space by exploiting the correlations among image keywords into a smaller number of higher level semantic topics. To tackle the second problem, we need to be able to reduce the bias between visual similarity and semantic similarity. We propose an iterative solution based on alternate optimization to achieve "optimal margins" jointly in both semantic and visual space. To achieve these aims, we present in this paper a new image annotation approach based on weighted KNN multi-label classification (KNNMLC).

The workflow of KNNLNC is as follows. For each training image, we first define a neighborhood under the visual similar measurement. Next a group of semantic topics are generated by using maximum margin clustering in the corresponding local keyword subspace of the defined neighborhood. By regarding each semantic topic as a class, SVM classification can be applied to learn the hyperplane to discriminate these semantic topics in the visual space. However instead of using the maximum margin approaches separately in the semantic space and the visual space, we propose using the alternate optimization approaches [4, 28] to achieve "optimal margins" in both spaces by maximizing the sum of the two margins to reduce the bias between the visual similarity and semantic similarity. Finally, we provide an annotation framework based on the weighted K Nearest Neighbor Classification (KNNC). The unlabeled image will be classified to the appropriate semantic topic correlated with each set of neighbor training images by the classifier. The keywords are then estimated and generated from the obtained topics and propagated to the unlabeled images. The experimental results on the ECCV2002 benchmark show that our method outperforms the state-of-the-art generative-based annotation method MBRM and discriminative-based model ASVM-MIL by 9% and 11% in $F1$ measure respectively.

The contributions of our approach are as follows:

- We propose to establish the correlations between semantic concepts and low-level features from a new perspective by adaptively modelling a local indicator function to explore the potential set of semantic contents correlated with each training image.

- We exploit the keyword contextual correlation information to generate semantic topics by using maximum margin clustering to deal with the multi-labeling problem.

- We develop an iterative solution to achieve "optimal margins" in both of the semantic feature space and the visual feature space in order to reduce the semantic gap problem.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 presents the annotation framework, and introduces the method of semantic topic generation and alternate optimization approach to achieve optimal margin in both spaces. We also present the technique for generating keywords from the topics and the annotation algorithm. The experimental results are presented in Section 4, with the conclusion given in Section 5.

2. RELATED WORK

A significant amount of research has been done to address the problem of AIA [34, 13, 17, 9, 16, 15, 19, 8, 30, 1, 23, 10, 18, 20, 12, 21, 11, 2, 31, 27, 26]. We will give a brief review of this line of research in this Section.

Starting from a training set of annotated images, many statistical learning models have been proposed to associate visual features with semantic concepts (keywords). Mori et al. [22] proposed a co-occurrence model, in which they formulated the co-occurrence relationships between keywords and sub-blocks of images. Duygulu et al. [7] generated the image visual words (blobs) vocabulary by clustering and discretizing the region features. They then utilized a machine translation model to form the links between the words and blobs to annotate new images. Latent semantic analysis (LSA) [12] and probabilistic latent semantic analysis (pLSA) [21] introduced latent variables to link image features with keywords. Florent et al. [21] built a linked pair of pLSA models to attach more importance to textual features. Jeon et al. [13] introduced a cross-media relevance model (CMRM), in which the region features are represented by discrete blobs. The CMRM modeling was subsequently improved through a continuous relevance model [17] and a multiple Bernoulli relevance model [9] respectively. Liu et al. [19] proposed a dual cross-media relevance model (DCMRM), which estimates the joint probability by the expectation over words in a pre-defined lexicon. Zhang et al. proposed a multi-label learning approach named ML-kNN, which is derived from the traditional k-Nearest Neighbor (kNN) algorithm [20]. Since the above methods did not explicitly treat semantic labels as image classes, they followed an unsupervised learning framework and the performance of annotation is strongly influenced by the quality of unsupervised learning and suffers from the problem of "semantic gap" [34]. For example, in discrete feature models, regions with different semantic concepts but similar appearance may

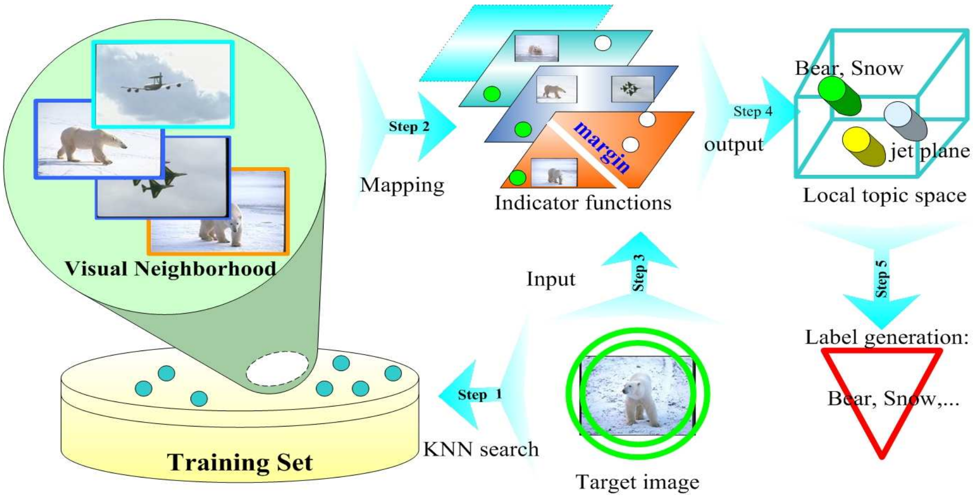


Figure 1: The framework of the proposed method

be clustered together to generate unreasonable visual keyword. The similar happens in continuous relevant model where the performance may be impaired by having training images with high visual generative probability but different semantics with the unlabeled images.

By viewing each keyword as an independent class, classification techniques have been applied to image annotation task [34, 11, 27, 5, 3, 8, 26] by building the different classification model for each keyword. Model-based methods [27, 5, 3] and SVM-based approaches [34, 11, 26] have both been applied to image annotation. Bernard et al. [3] proposed a gaussian mixture models for each semantic concepts and employed EM algorithm for parameter learning. Shi et al. [27] proposed a Bayesian learning framework of hierarchical multinomial mixture models of concepts for image annotation. Gao et al. [11] introduced a multi-resolution grid-based annotation framework for image content representation and a hierarchical boosting algorithm to address the problem using classification technique. Yang et al. [34] proposed a asymmetrical support vector for region-based image annotation. Carneiro et al. [5] formulated the image annotation as supervised multi-class problem, and learned the distribution model for each class. In their methods, the discriminative information between different classes is not explicitly exploited. Fan et al. [8] proposed a hierarchical classification framework for bridging the semantic gap and achieving multi-level image annotation automatically.

Recently, the multi-labeling problem in AIA has attracted more research attentions. Jin et al. [14] exploited EM algorithm to fit a coherent language model to generate an annotation keyword subset. Kang et al. [16] proposed a correlated label propagation method for multi-label learning, in which they explicitly exploited high-order correlations between semantic labels based on properties of submodular functions. Zhou et al. [36] made use of keyword correlations to perform concept (keyword subset) based annotation. Qi et al. [25] presented a correlative multi-label (CML) annotation framework which simultaneously classified concepts and modelled their correlations between them in a single step.

3. OUR METHOD

This section presents our framework for image annotation. We will show that our method establishes the new correlations between concepts and low-level features by capturing the correlations between different concepts and the discrimination between visually similar concepts.

3.1 The Annotation Framework

Let $T = \{\{f_1, a_1\}, \{f_2, a_2\}, \ldots, \{f_{|T|}, a_{|T|}\}\}$ denote the set of labeled examples, where T is the size of the training set. f_i and a_i are respectively the visual features and annotation keywords of image i. Each annotation can also be represented as a vector: $z_i = \{z_{i1}, z_{i2}, \ldots, z_{i|V|}\}$, where V denote the keyword vocabulary, and $|V|$ is the size of the V. If the kth keyword belongs to the annotation set a_i, then z_{ik} is set 1; else z_{ik} is 0. Given a new image $I_u = \{f_u\}$, our goal is to find its annotation a_u.

Due to the problem of "semantic gap", for a given training image J, it is often the case that the images in J's visual similar neighborhood may not share any semantic keywords with J or are obviously belonging to multiple different semantic classes. Thus, for each training image J_i and unlabeled image I_u, we define two vectors $\mathbf{I}^i$ and $\mathbf{P}^i$ as follows:

$$\mathbf{I}^i = \begin{bmatrix} I(g_i(I_u) = 1) \\ \cdots \\ I(g_i(I_u) = k) \\ \cdots \\ I(g_i(I_u) = s) \end{bmatrix} \quad \mathbf{P}^i = \begin{bmatrix} P_1(v|J_i) \\ \cdots \\ P_k(v|J_i) \\ \cdots \\ P_s(v|J_i) \end{bmatrix}, \tag{1}$$

where s is the number of the different semantic classes implied by J_i. Function $g_i(I_u)$ outputs the object class that the unlabeled image belongs to. $I(g_i(I_u) = k)$ is an indicator function that outputs 1 when the value of object class $g_i(I_u)$ is k and zero otherwise. $P_k(v|J_i)$ represents the probability density of keyword v being generated from the kth semantic class distribution. Then in our method, the score of assigning keyword v to the unlabeled image I_u could be estimated based on weighted K nearest neighbor classification, which is given by:

$$s(v) = \Sigma_{J_i \in KNN(I)}\ w_i \ < \mathbf{I}^i \cdot \mathbf{P}^i > \tag{2}$$

where $< x \cdot y >$ denotes the inner product between x and y.

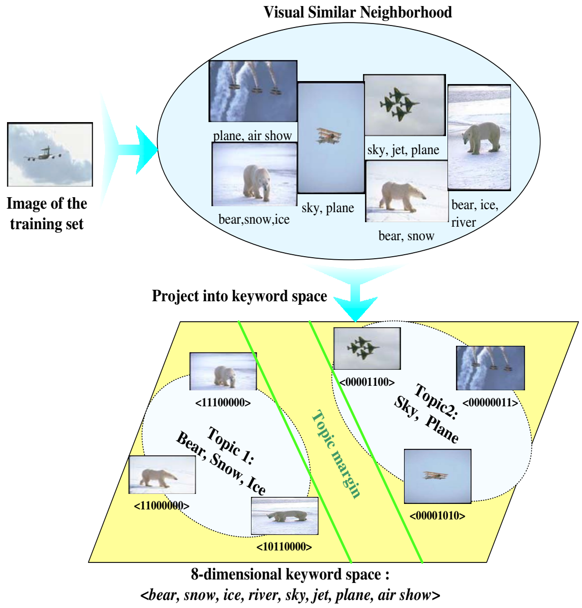


Figure 2: The maximum margin clustering based semantic topic generation. Images contained co-occurred keywords are likely to be assigned in the same topic. Different topics have large "distance" or "margin" in the semantic space.

w_i serves as the weight of each training image, consisting of two parts: the visual similarity between I_u and J_i; and the visual discriminative power of J_i, which will be introduced in Section. 3.5.

Figure 1 illustrates the framework of our annotation method. Instead of simply propagating the annotation of the visually similar training images to the unlabeled images, with the use of indicator function, our method attempts to exploit the underlying semantics implied by the visual similarity between the unlabeled image and its neighbors. Inherently, there could be a 1 to m mapping in $\mathbf{I}$. For simplicity we limit only to 1 to 1 mapping in $\mathbf{I}$ in this paper. Thus our framework can be rewritten as:

$$s(v) = \quad \Sigma_{J_i \in KNN(I)} \; w_i \; P_{g_i(I_u)}(v|J_i), \tag{3}$$

where the generative probability distribution of the semantic keywords of the unlabeled image is determined by the unique value of $g_i(I_u)$. Thus $g_i(I_u)$ is served as the *indictor function*, which is used to differentiate semantic classes with similar visual appearance to J_i and classify the unlabeled image to the relevant ones.

In next Section, we will discuss how to establish the appropriate "semantic classes" and how to differentiate them in the visual space.

3.2 Maximum Margin Clustering based Semantic Topic Generation

Being different from the traditional multi-class classification problem, image annotation always leads to a complex multi-label classification problem because each image may potentially correspond to multiple correlated labels [16, 25]. For example, Figure 2 shows a training image and its visual similar neighborhood. From the Figure, we can observe that the number of different keywords in $N(J)$ is more than the number of the images containing the same keywords. Moreover, the training images belonging to different keywords are partly overlapping. Thus we can imagine that in a large scale concept space, it is difficult to train classifier for each keyword [5]. On the other hand, from the global perspective, we can observe that most of the images can be viewed as a kind of description of some events/topics of the real world. Thus, to a certain extent we can classify them according to high level topics.

There are many research efforts on topic generation and detection in text retrieval [29, 6]. A straightforward approach is to group images into different *subsets* according to the keyword correlations between their semantic labels. Images within the same subset will be considered as belonging to the same topic. Thus we can regard each topic as a class, and each image associated with a class label y.

Let $N(J) = \{x_1, \ldots, x_k\}$ denote the top k neighbor of the training image J, where x_i is represented as a keyword vector $\{z_{i1}, \ldots, z_{i|V|}\}$. Let each keyword correspond to one dimension in the semantic (concept) space, the keywords in $N(J)$ determine a *local keyword subspace.* The topical similarity between image x_i and x_j in this space can be measured by their inner product $< x_i \cdot x_j >$, which is more generally denoted by $K_s(x_i, x_j)$. Then images containing co-occurring keywords always have large semantic similarity. Figure 2 shows 6 images distributed in a 8-dimensional local keyword subspace. Each image is represented with a 8 dimensional vector. Different topics have large "distance" or "margin" in the semantic space. We can use the maximum margin clustering [32, 33] to obtain 2 groups of images with maximum margin between them. Here we use support vector regression (SVR) based maximum margin clustering [35] to obtain our semantic topics with maximum semantic margins, that is:

$$\begin{aligned} \min_{y, w_s, b_s, \xi, \xi^*} \quad & ||w_s||^2 + 2C(\xi + \xi^*)^T e, \\ s.t. \quad & y_i - (w_s^T \phi(z_i) + b_s) \leq \xi_i, \\ & -y_i + (w_s^T \phi(z_i) + b_s) \leq \xi_i^* \\ & \xi_i \geq 0, \xi_i^* \geq 0, \\ & -l \leq e^T y \leq l. \end{aligned} \tag{4}$$

where C, ξ and ξ^* are the same as in the standard supervise learning SVR. Specifically, ξ_i and ξ_i^* are the slack variables for the errors, $C > 0$ is a regularization parameter and e is the vector of ones. ϕ is the mapping function induced by a kernel, where we have $\phi(z_i) \cdot \phi(z_j) = K_s(z_i, z_j)$. l is the balance parameter which prevents the trivially "optimal" solution where $x_i's$ are assigned to the same class and the resultant margin will be infinite.

3.3 Maximizing the Sum of the Margins of Semantic Space and Visual Space

After obtaining the semantic topics and regarding each topic as a class, we need to find out which topic is relevant to the unlabeled image I_u according to the visual features of the topics and the unlabeled image. Obviously, it is a kind of supervised classification and seems that the maximum margin classification approach such as SVM classification can be applied directly. However, the situation is more complicated, because of the mismatch that may occur between the margins of the semantic feature space and the visual fea-

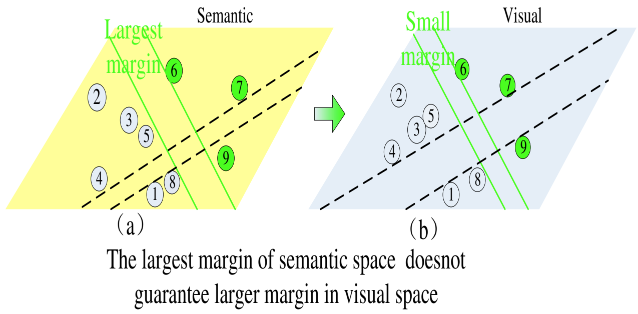

Semantic
Visual
Largest margin
Small margin
1
2
3
4
5
6
7
8
9
(a)
(b)
The largest margin of semantic space doesnot guarantee larger margin in visual space

Figure 3: Using the maximum margin techniques separately in semantic and visual space

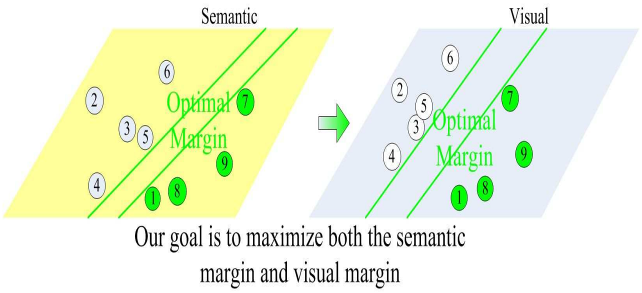

Semantic
Visual
Optimal Margin
Optimal Margin
1
2
3
4
5
6
7
8
9
Our goal is to maximize both the semantic margin and visual margin

Figure 4: Maximize the margins in both semantic and visual space

ture space. Specifically, the "optimal" topics obtained by the maximum margin clustering do not always lead to a good visual margin given by the SVM classification in the visual space.

We illustrate this in Figure 3 and Figure 4. In both Figures, we have 9 images projected into the semantic feature space and visual feature space respectively. For simplicity, Figure 3 (a) only shows two candidate margins derived in the semantic space: the larger green one and the smaller black one, each of them can be used to cluster the images into different sub clusters (topics). With the maximum margin clustering procedure, the green margin will be selected to cluster the images into two topics, where *topic*1 is consisted of green points, the *topic*2 is consisted of gray points. We name the green margin as the *largest green margin* and the alternate black margin as the *small black margin*. After clustering, the topics will be assigned to the images as class labels. Next SVM classification procedure will be applied to seek the maximum margin in the visual space to separate the images according to their corresponding class labels. Therefore, it can be easily observed that different class label assignments may result in different visual maximum margins learned by SVM classification. Figure 3 (b) shows that with the topics obtained by the *largest green margin* in the semantic space, the corresponding visual margin learned by SVM classification is smaller than that of the alternate *small black margin*. According to the SVM theory, larger margin means better generation ability, thus in order to get the overall optimal performance, we should achieve optimal margin in both of the semantic space and feature space. To the end, Figure 4 shows that if we choose a proper margin to cluster the topics, better margins can be achieved in both of spaces.

In this paper, we propose to maximize the sum of the two margins by forcing the semantic topic label y to be consistent in both spaces. The goal and constrains are defined as follows:

$$\begin{aligned}
\min_{y,w_s,w_f,b_s,b_f,\xi,\xi^*,\xi'} \quad & ||w_s||^2 + ||w_f||^2 + 2C(\xi+\xi^*)^T e + 2C'\xi'^T e \\
s.t. \quad & y_i - (w_s^T\phi(z_i)+b_s) \le \xi_i, \\
& -y_i + (w_s^T\phi(z_i)+b_s) \le \xi_i^* \\
& y_i(w_f^T\phi'(f_i)+b_f) \ge 1-\xi_i' \\
& \xi_i \ge 0, \xi_i^* \ge 0, \\
& \xi_i' \ge 0, \\
& -l \le e^T y \le l.
\end{aligned} \tag{5}$$

where ϕ' is the mapping function induced by a kernel, we define $\phi'(f_i)\cdot\phi'(f_j) = K_f(f_i,f_j)$, which measures the distance (similarity) of x_i and x_j in the visual space.

However, in optimization theory, maximizing two margins simultaneously is a hard problem. Recently, there are some work using alternate optimization [4] technique to deal with the problem of maximum margin learning. Here we just name a few related work. Zhang et al. [35] proposed using alternate optimization and SVR to deal with the problem of maximum margin clustering; While Steven et al. [28] applied alternate optimization to solve the Coupled SVM (CSVM) classification. In this work, we propose to exploit alternate optimization to solve Eqn. 5. In CSVM, the class labels of the labeled data are pre-determined, while that in our framework are more "unsupervised", in which the class labels are learned by maximizing the sum of the margins of SVM clustering and classification.

We provide an iterative procedure for solving Eqn. 5. First we run a simple clustering method such as k-means to obtain the class label y and try to find the $< w_s, b_s >$ and $< w_f, b_f >$ that optimize the objective function. By fixing y, in accordance with Eqn. 5, we then try to solve the following two optimization problems:

$$\begin{aligned}
\min_{w_s,b_s,\xi,\xi^*} \quad & ||w_s||^2 + 2C(\xi+\xi^*)^T e, \\
s.t. \quad & y_i - (w_s^T\phi(z_i)+b_s) \le \xi_i, \\
& -y_i + (w_s^T\phi(z_i)+b_s) \le \xi_i^* \\
& \xi_i \ge 0, \xi_i^* \ge 0, \\
& -l \le e^T y \le l. \\
and \quad & \\
\min_{w_f,b_f,\xi'} \quad & ||w_f||^2 + C\xi'^T e \\
\text{s.t.} \quad & y_i(w_f^T\phi'(f_i)+b_f) \ge 1-\xi_i', \\
& \xi_i' \ge 0
\end{aligned} \tag{6}$$

Second, when $< w_s, b_s >, < w_f, b_f >$ are solved from the first step, we can fix them and turn to finding the optimal y that fits the data. With w_s, w_f at hand, the original optimization problem can be simplified into the following form:

$$\begin{aligned} min_{y,\xi,\xi^*,\xi'} \quad & 2C(\xi+\xi^*)^T e + 2C'\xi'^T e \\ s.t. \quad & y_i - (w_s^T \phi(z_i) + b_s) \le \xi_i, \\ & -y_i + (w_s^T \phi(z_i) + b_s) \le \xi_i^* \\ & y_i(w_f^T \phi'(f_i) + b_f) \ge 1 - \xi_i' \\ & \xi_i \ge 0, \\ & \xi_i^* \ge 0, \\ & \xi_i' \ge 0, \\ & -l \le e^T y \le l. \end{aligned} \quad (7)$$

This can be reduced to:

$$min_y \Sigma_{i=1}^K \begin{cases} C|w_s^T \phi(z_i) + b_s - y_i| + \\ C' max(0, 1 - y_i(w_f^T \phi'(f_i) + b_f)) \end{cases} \quad (8)$$

Since we have an additional constraint that y must be integer, the above optimization is an integer programming problem. If there are many variables, integer programming can be a very hard problem (NP-complete). Fortunately, there is a small K in this case and thus we can solve our problem reasonably efficiently. The detailed solution is listed in Algorithm 1.

Algorithm 1 : maximum margins in semantic and visual space

```
1: Input: The neighborhood $N(J)$ for training image $J$.
2: Initialize the labels $y$ by using simple clustering method
   such as the k-means in the semantic space.
3: Compute $w_s, b_s$ and $w_f, b_f$ from Eqn. 6.
4: Update the labels according to Eqn. 8
5: Repeat steps 2-5 until convergence.
6: Return topic label $y$, and parameters $w_s, b_s$, $w_f, b_f$.
```

After obtaining the classification vector $< w_f, b_f >$, we can classify the unlabeled image to the appropriate semantic topic:

$$g(I_u) = y_u = h(w_f^T(f_u) + b_f), \quad (9)$$

where $h(\cdot)$ is a function of mapping the value of $w_f^T(f_u) + b_f$ into $g(I_u)$. For example, for two class classification problem, $h(\cdot)$ is a simple 0-1 indictor function. When $w_f^T(f_u) + b_f > 0$, it outputs a 1; otherwise 0. With the optimal margins in both semantic space and visual space, we can achieve strong ability to discriminate different semantic topics with visual similarity and classify the unlabeled image to the relevant topic.

3.4 Keyword Distribution Estimation on the Semantic Topic

It should be noted that the topic derived from the SVM clustering is related to the keyword distribution generated from the annotations of the images contained in the corresponding cluster (topic). Therefore the keyword generative probability distribution $P_k(v|J)$ associated with topic k can be estimated as follows:

$$P_k(v|J) = \frac{\sum_{x:y_x=k} \delta_v K_f(x,J)}{\sum_{x:y_x=k} \sum_{v'} \delta_{v'} K_f(x,J)}, \quad (10)$$

where if v is the annotation of x, then δ_v is set to 1, otherwise 0. $K_f(x,J)$ measures the visual similarity between x and J. It can be regarded as a weight for adjusting the contributions of images to the generative probability distribution of keywords of the given topic. In other words, the nearer neighbors will give more contributions to the probability estimation of the keyword generative distribution.

3.5 Estimating the Weight of the KNN Classification

Our framework is based on the weighted KNN classification. Thus the weight estimation is an important issue. In this Section we will present a novel method for weight estimation. Recall that most previous works employ the distance (similarity) between neighbor J_i and unlabeled image I_u as the weight, meaning that nearer neighbors will weight more than farer ones. In addition to considering this kind of similarity between J_i and I_u, we also estimate the *visual discriminative power (VDP)* of the training images to weight the contribution of J_i. The VDP is defined as the ability of the training data to differentiate itself from the semantically irrelevant images under the visual similarity measurement. An image with weak VDP means that there are more semantically different images occurred in its visual neighborhood. Therefore, we adopt the heuristic that weak VDP images should give smaller contribution in image annotation, and vice versa. Thus, in Eqn. 2, the weight w_i can be defined as follows:

$$w_i = K_f(I_u, J_i) * VDP(J_i). \quad (11)$$

For a given training image I, we can obtain $VDP(I)$ by counting the numbers of labels of the top-ranked images being consistent with the annotations of I. The larger the value, the larger the VDP of the training data, which is defined by:

$$VDP_N(I) = \sum_{J_i \in KNN(I)} \beta_i \, a_I \cap a_{J_i}, \quad (12)$$

where β_i is the weight coefficient, which measures the contribution of each J_i to the VDP of I. The more visual similar with I, the more contribution is given by J_i. $a_I \cap a_{J_i}$ measures the semantic agreement between I and J_i.

Next, we consider another image set $IN(I)$, termed as "*Inverted Neighborhood*", defined by:

$$IN(I) = \{J | I \in N(J)\}. \quad (13)$$

$IN(I)$ is composed of images whose visual similar neighborhood contain image I. Since the feature similarity measurement is not always symmetrical, it means that $IN(I)$ and $N(I)$ may be different. If an image has large VDP, the semantics keywords correlated with $IN(I)$ and I are expected to be similar. To examine the semantic coincidence between I and $IN(I)$, we can define the VDP estimation based on $IN(I)$ by following the same way as Eqn. 12:

$$VDP_{IN}(I) = \sum_{J_i \in IN(I)} \beta_i \, w_I \cap w_{J_i}, \quad (14)$$

where β_i plays the same role as in Eqn. 12. Obviously, if I appears in the top position of the visual similar neighborhood of J_i, then J_i will have a large contribution to the VDP of I.

By combining the above estimation with a fixed small constant c, we obtain the final definition of VDP of image I as:

$$VDP(I) = \lambda_1 \, VDP_N(I) + \lambda_2 \, VDP_{IN}(I) + \lambda_3 c, \quad (15)$$

where $\lambda_3 = 1 - \lambda_1 - \lambda_2$. c is the fixed constant value smoothing the estimation.

3.6 The Annotation Algorithm

We now present our weighted KNN multi-label classification based annotation algorithm. Our annotation process is consisted of two parts. The first part (steps 2 - 5) is the preprocessing procedure of local multi-label classification, which obtains the value of the indicator function and keyword distributions of the topics of each training image. The second part (steps 6 - 13) is based on the result of the first preprocessing procedure to label the unlabeled image in a weighted KNN classification framework.

The annotation algorithm is given as follows:

```
Algorithm 2 : Anno_KNNMLC
1: Input: unlabeled image $I_u$, words vocabulary $V$, training set $T$, fixed annotation size $m$
2: for i = 1, 2, ..., |T| do
3:     learn the semantic topic distribution $P_k(\cdot|J_i)$ and the corresponding classification function $g_i(x)$ based on $N(J_i)$ according to Algorithm 1.
4:     Learn the visual discriminative power $VDP(J_i)$ according to Eqn. 15.
5: end for
6: Given the unlabeled image $I_u$, choose the neighborhood of $I$
7: for i = 1, 2, ..., k do
8:     Classify $I_u$ to the appropriate semantic topic according to $g_i(I_u)$.
9:     Obtain $I^i$ and $P^i$ according to Eqn. 1.
10:    Calculate the weight $w_i$ for each $J_i$ according to Eqn. 11.
11: end for
12: Calculate the ranking score $s(v)$ according to Eqn. 2.
13: Output: choose $m$ keywords with largest $s(v)$ as the annotation of $I_u$.
```

It can be concluded from Algorithm 2 that the proposed method has the following three advantages. 1) It has scalability to a large scale concepts of interest since we follow a KNN classification framework. 2) It has strong discriminative power because instead of taking each keyword as a class, we discriminate the images at a higher semantic level, i.e. the topic level. The maximum margin in both semantic space and visual space also increases the discriminative ability of our method. 3) It is efficient, because both semantic clustering and visual classification are implemented in small neighborhood.

4. EXPERIMENT

4.1 Setup

We tested our algorithm using Corel data set provided in [7]. The data set consists of 5000 images from 50 Corel Stock Photo CDs. Each image is linked with 1-5 annotation words. There is a total of 374 different words in the data set. The data set is divided into two parts: 4,500 images for training and rest of 500 for test. Every image in the data set is partitioned into rectangular grids, and a feature vector is calculated for every grid region (grid-based features). The number of rectangles is 17, which includes 16 regular grids and one extra grid in the center of the image. The feature vector is 528 dimension: 448 color features (including local and global color histogram) and 80 edge features extracted according to MPEG7. In the following experiments, the size of the neighborhood is set to 25. We use the Gaussian kernel as our kernel function K_s. For kernel K_f, we use non-parametric Gaussian kernel based density estimation to estimate the generative probability between each training pair, and convert it as a predefined kernel.

Similar to previous studies for image annotation, we use recall, precision and F_1 to measure the effectiveness of the algorithm. Given the query word w, if there are $|W_G|$ human annotated images with label w in the test set, $|W_M|$ annotated images with this label by our method, where $|W_C|$ are correct, then the recall, precision and F_1 are defined as:

$$Recall = \frac{|W_C|}{|W_G|}, \qquad Precision = \frac{|W_C|}{|W_M|}, \tag{16}$$

$$F_1 = \frac{2 * Recall * Precision}{Recall + Precision} \tag{17}$$

.

4.2 Experimental Results

4.2.1 Comparison with Relevance Model

Figure 5 shows the results of the proposed method (KNN-MLC) on the test set, where the recall and precision are averaged over 263 words. We also present the results of other related work under the same experiment setting. Specifically, we consider: TM [7], CMRM [13], and MBRM [9]. For all the methods, the annotation size for each image is set to 5. In MBRM, non-parametric Gaussian kernel estimation is exploited to estimate the feature generative probability distribution.

From Figure 5, it is clear that the continuous feature estimation (MBRM) is better than discrete feature estimation (TM & CMRM), because it increases the accuracy of visual similar measure by non-parametric kernel estimation. So in the following comparisons, we will only focus on MBRM. Figure 5 shows that the proposed method achieves the best performance in the comparison. Compared with MBRM, it gains 13%, 5% and 9% on recall, precision and F_1 respectively. It shows that our method is able to achieve encouraging improvements on all three measures, especially the average recall is greatly improved. This demonstrates that the proposed framework is effective to discriminate the different semantic topics for the training data, and hence increase the possibility of propagating the correct keywords to the unlabeled images.

4.2.2 Comparison with Discriminative Model

In order to further evaluate the effectiveness of our method, we also compare our method with ASVM-MIL [34]. ASVM-MIL treats AIA as the problem of multiple-instance learning and proposes an asymmetrical support vector machine to address the problem. In order to be comparable, we report the experimental results on 70 frequent concepts, which are the same as the test environment [34] used in ASVM-MIL. We also use the same visual features for images as [34] in this experiment. Specifically, the images are segmented into regions by using the Normalization Cut, before extracting a 36-dimensional feature from each region (region-based features), which includes color, texture and area features. The

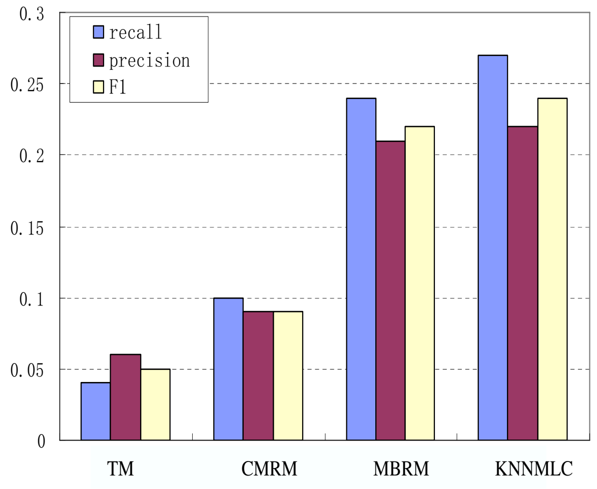


Figure 5: The effectiveness of our method compared with TM, CMRM and MBRM

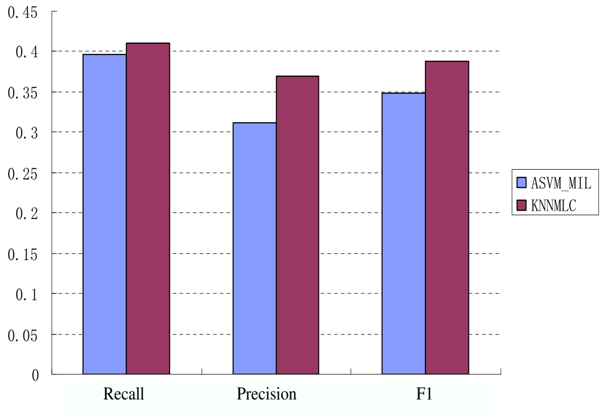


Figure 6: The effectiveness of our method compared with ASVM-MIL

results in Figure 6 demonstrate that the average recall and precision measures of our method are significantly higher than ASVM-MIL, with the F_1 measure improved by 11% from 0.349 to 0.388. This is because we exploit the keyword contextual correlation to obtain the semantic topics, which reduce the number of classes and the scale of multi-labelling problem, leading to better discriminative power.

4.2.3 *The Effectiveness of Visual Discriminative Power Estimation*

In order to evaluate the effectiveness of the proposed visual discriminative power estimation, we convert the obtained visual discriminative power into prior probability and apply it to MBRM for comparison purpose, which assumes that each training image has uniformly prior probability distribution. We report the result by exploiting the grid and region based features respectively in Figure 7. From the Figure, it can be observed that the method uses the prior estimation always perform better than the corresponding ones with uniformly prior probability. The reason is that we assign a small prior probability to the training image with weak visual discrimination, and then reduce their contribution in the annotation. Meanwhile, the contributions of the strong visual discriminative ones are increased. This process enhances the probability of propagating the correct semantic labels to unlabeled images, which brings more improvement to the performance. From Figures 5 and 7, we also observe that the annotation performance of KNNLCM is higher than when VDP estimation is used. This demonstrates that the proposed KNNLMC gains benefits from two side: the local multi-label classification and the VDP estimation.

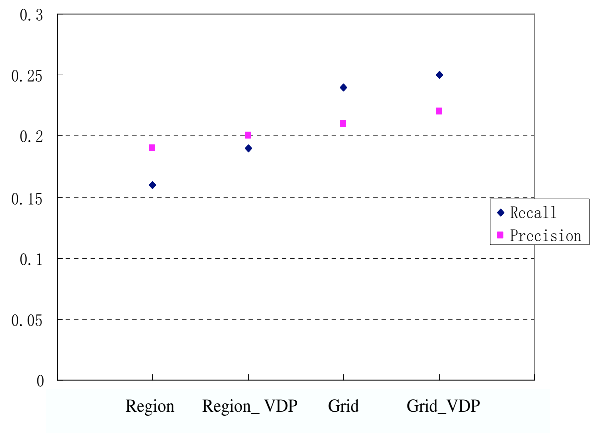


Figure 7: The effectiveness of the visual discriminative power estimation

4.2.4 *Keyword Correlation Comparison*

As we have mentioned earlier, in our method, each semantic topic is obtained from the semantic cluster. Since each semantic cluster consists of semantically similar images, the semantic keywords in each topic are relatively correlated. Therefore, the annotation generated from the proposed framework are expected to implicitly incorporates the keyword correlation information. In order to verify the effectiveness of the proposed method, the experimental results for two-word queries are shown in Table 8, where the queries are based on the most co-occurred pairs of keywords. It is obvious that the average recall and precision are improved significantly as compared to MBRM where keywords are annotated independently. Take the most co-occurred keyword pair “sky, tree” as an example, the recall and precision are improved from 34% and 8% to 42% and 9% respectively as compared to MBRM. The results in Figure 8 clearly indicate that the proposed topic generation process provides an effective way to exploit keyword correlation in the annotation process.

4.2.5 *Some Annotation Examples*

To further illustrate the effectiveness of our method for AIA, we lists some images labeled by our method in Figure 9, together with the manual annotations and the annotations generated by MBRM. From Figure 9, we observe that, first, the annotations generated from our method are more semantically related to each other than those from MBRM. Take the first image as an example, “swimmer, pool, water” should be co-occur more together than “water, mountain, nest”. Since our method uses maximum margin clustering in concept space to generate the local topic, which brings

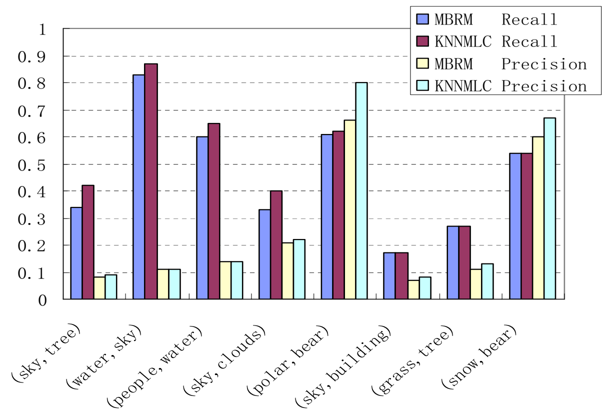


Figure 8: Results for two-word queries, evaluated on some examples of most co-occurred pairs of keywords

concept contextual correlation into our annotation process and leads to more accurate annotations. Second, MBRM is easily influenced by images with similar appearance but different semantics. For example, image 4, 6, 10 have wrong annotation labels according to MBRM model. While our method can generate more correct annotation. This is because the use of maximum margin in both semantic space and visual space helps us to find the correct topics for the unlabeled image, and thus improves the overall annotation performance.

5. CONCLUSIONS

In this paper, we propose a novel annotation method which establishes the correlations between semantic concepts and low-level features from a new perspective. By adaptively modelling a local multi-label classification indictor function, our method captures the keyword contextual correlations and also exploits the discrimination between visual similar concepts. A series of maximum margin techniques (maximum margin clustering and maximum margin classification) are performed on a training neighborhood of the unlabeled image to obtain the value of the indicator function, which differentiates semantic classes with similar visual appearance and reduces the bias between visual similarity and semantic similarity. The experimental results demonstrate the effectiveness of our method.

In our current work, we assume that each unlabeled image corresponds to one topic implied by the training image. In our future work, we plan to improve the current classification method with a 0-1 posterior probability in order to establish the one to many mapping. In our framework, it is a time-consuming and labor-intensive process to find the neighborhood of the images. In our future work, we also plan to exploit the index structure to speed up this process. Finally, we need to exploit more information such as the web information to enhance our estimation for the underlying semantics of visual similarity.

6. REFERENCES

[1] A. J. A. Vailaya and H. Zhang. On image classification: City vs. landscape. *Pattern Recognition*, pages 1921–1936, 1998.

[2] A.Yavlinsky, E.Schofield, and S.Ruger. Annotation using global features and robust nonparametric density estimation. *CIVR*, 2005.

[3] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. *ICCV*, pages 408–415, 2001.

[4] J. Bezdek and R. Hathaway. Convergence of alternating optimization. *Neural, Parallel Sci. Comput.*, 11(4):351–368, 2003.

[5] G. Carneiro, A. Chan, and N. V. P.J. Moreno. Supervised learning of semantic classes for image annotation and retrieval. *PAMI*, 2007.

[6] S. Deerwester, S. Dumais, G. Furnas, L. T. K., and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, pages 391–407, 1990.

[7] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *ECCV*, 2002.

[8] J. Fan, Y. Gao, and H. Luo. Hierarchical classification for automatic image annotation. *SIGIR*, pages 111–118, 2007.

[9] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. *CVPR*, pages 1002–1009, 2004.

[10] D. Forsyth and M. Fleck. Body plans. *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, pages 678–683, 1997.

[11] Y. Gao, J. Fan, X. Xue, and R. Jain. Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. *ACM Multimedia*, pages 901–910, 2006.

[12] F. D. Gatica-Perez and R. Oka. On image auto-annotation with latent space models. *ACM Multimedia*, pages 275–278, 2003.

[13] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. *SIGIR*, pages 119–126, 2003.

[14] R. Jin, J. Y. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. *ACM Multimedia*, pages 892–899, 2004.

[15] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & wordnet. *ACM Multimedia*, pages 706–715, 2005.

[16] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. *CVPR*, pages 291–294, 1719-1726 2006.

[17] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. *NIPS*, pages 553–560, 2004.

[18] Y. Li and L. Shapiro. Consistent line clusters for building recognition in cbir. *Proc. Intąŕl Conf. Pattern Recognition*, pages 952–956, 2002.

[19] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. *ACM Multimedia*, pages 605–614, 2007.

[20] Z. Z. M.L. Zhang. Ml-knn: A lazy learning approach to multi-label learning. *pattern recognition*, pages 605–614, 2007.

[21] F. Monay and D. GaticaPerez. Plsa-based image autoannotation: Constraining the latent space. *ACM*

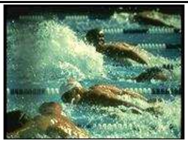
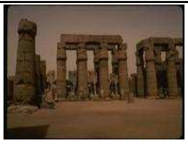
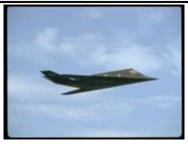
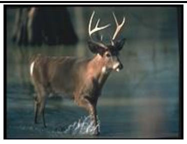
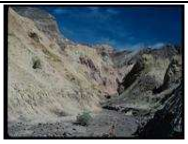
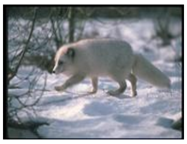
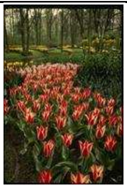
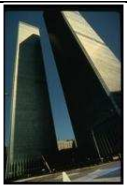
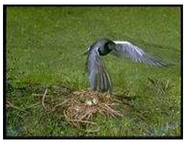
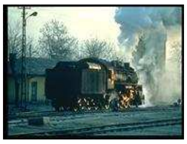
					
MBRM	people, water, mountain, nest, tree	people, statue, sky, water, tree	plane, sky, people, tracks, cars	water, people, sky, tree, grass	mountain, sky, tree, water, bridge
KNNLMC	people, pool, swimmer, water, tree	pillar, statue, stone, people, tree	jet, plane, sky, smoke, prop	deer, white-tailed, water, tree, sky	valley, sky, mountain, tree, water
Ground Truth	people, pool, swimmer, water	pillar, sculpture, statue, stone	jet, plane, sky	deer, white-tailed, river, water	valley, sky, plants, desert
					
MBRM	snow, nest, birds, wood, cliff	leaf, woman, garden, tulip, people	hats, castle, close-up, building, sky	plants, leaf, flowers, water, close-up	water, horses, statue, people, tree
KNNLMC	fox, snow, arctic, rocks, grass	flower, garden, tulip, leaf, close-up	tower, building, city, sky, clouds	nest, birds, flowers, water, leaf	locomotive, railroad, train, snow, water
Ground Truth	fox, snow, arctic	flower, garden, tree, tulip	tower, street, building	birds, nest, fly	locomotive, train, building, smoke

Figure 9: Comparison of the annotation results of ten sample images provided by MBRM and KNNLMC.

Multimedia, pages 348–351, 2004.

[22] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. *In First International Workshop on Multimedia Intellegent Storage and Retrieval Management*, 1999.

[23] Z. M. N. Haering and N. Lobo. Locating dedicuous trees. *Proc. Workshop in Content-Based Access to Image and Video Libraries*, pages 18–25, 1997.

[24] M. Naphade, J. R. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, and A. Hauptmann. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13:86–91, Jul-Sep 2006.

[25] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H. Zhang. Correlative multi-label video annotation. *ACM Multimedia*, pages 17–26, 2007.

[26] X. Qi. and Y. Han. Incorporating multiple svms for automatic image annotation. *Pattern Recognition*, 40:728ÍC741, 2007.

[27] R.Shi, T. Chua, C. lee, and S. Gao. Bayesian learning of hierarchical multinomial mixture models of concepts for automatic image annotation. *CIVR*, pages 102–112, 2006.

[28] C. Steven, R. Michael, and R. Jin. Integrating user feedback log into relevance feedback by coupled svm for content-based image retrieval. *Data Engineering Workshops*, pages 1177–1186, 2005.

[29] B. Sun, P. Mitra, H. Zha, C. L. Giles, and J. Yen. Topic segmentation with shared topic detection and alignment of multiple documents. *SIGIR*, pages 199–206, 2007.

[30] M. Szummer and R. Picard. Indoor-outdoor image classification. *Workshop Content-Based Access to Image and Video Databases*, 1998.

[31] J. Tang and P. Lewis. Using multiple segmentations for image auto-annotation. *CIVR*, pages 581–586, 2007.

[32] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. *NIPS*, 2004.

[33] L. Xu and D. Schuurmans. Unsupervised and semi-supervised multi-class support vector machines. *NIPS*, 2006.

[34] C. Yang and M. Dong. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. *CVPR*, pages 2057–2063, 2006.

[35] K. Zhang, I. Tsang, and J. Kwok. Maximum margin clustering made practical. *ICML*, 2007.

[36] X. Zhou, M. Wang, Q. Zhang, J. Zhang, and B. Shi. Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. *CIVR*, pages 25–32, 2007.