Automatic Image Annotation By An Iterative Approach: Incorporating Keyword Correlations And Region Matching

Xiangdong Zhou1, Mei Wang1, Qi Zhang2, Junqi Zhang1, Baile Shi1
1 Fudan University, Shanghai, China
2 University of North Carolina at Chapel Hill, USA
{xdzhou, 051021052, 041021054, bshi}@fudan.edu.cn, zhangq@email.unc.edu

ABSTRACT
Automatic image annotation automatically labels image content with semantic keywords. For instance, the Relevance Model estimates the joint probability of the keyword and the image [3]. Most of the previous annotation methods assign keywords separately. Recently the correlation between annotated keywords has been used to improve image annotation. However, directly estimating the joint probability of a set of keywords and the unlabeled image is computationally prohibitive. To avoid the computation difficulty we propose a heuristic greedy iterative algorithm to estimate the probability of a keyword subset being the caption of an image. In our approach, the correlations between keywords are analyzed by “Automatic Local Analysis” of text information retrieval. In addition, a new image generation probability estimation method is proposed based on region matching. We demonstrate that our iterative annotation algorithm can incorporate the keyword correlations and the region matching approaches handily to improve the image annotation significantly. The experiments on the ECCV2002 benchmark show that our method outperforms the state-of-the-art continuous feature model MBRM with recall and precision improving 21% and 11% respectively.

Categories and Subject Descriptors
I.4.8 [Image Processing and Computer Vision]; H.3.3

General Terms
Algorithms, Experimentation, Performance

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands
Copyright 2007 ACM 978-1-59593-733-9/07/0007 ...$5.00.

This work was partially supported by the Natural Science Foundation of China Grant No.60403018, the Natural Science Foundation of Shanghai China Grant No.04ZR14011, Fundamental Research 973 Program of China Grant No.2005CB321905

Keywords
Image annotation, continuous feature model, words correlation

1. INTRODUCTION

Image semantic annotation – associating keywords or captions to the image, is the key step leading to the semantic keyword based image retrieval, which is considered to be convenient and easy for most ordinary users. The early annotation approaches rely on professionals or experts for annotation. It suffers from the problems of labor intensity and subjectivity. With the rapid growth of image archives, various automatic image annotation approaches based on machine learning and statistical models have been proposed [1, 2, 3, 4, 5, 9, 10, 12, 13, 14, 15, 16, 17, 21, 23, 24, 27]. However, due to the well known “semantic gap” problem, the performance of image auto-annotation is still urgent to be improved.

The basic idea shared by most previous work is that the visual features belonging to the same keyword are coherent. The image segmentation approach is usually applied to obtain the image regions which contain certain semantic meanings in the ideal conditions. Therefore, the images or image regions with similar visual features can be grouped together and associated with a certain set of keywords. The discrete feature models such as Translation Machine model [2] and Relevance Model CMRM [3] generate the image visual words vocabulary by clustering and discretizing the region features, and then the correlation between the visual words and semantic keywords are estimated. Continuous feature model such as CRM [5] and MBRM [4] use non-parameter kernel-based density estimation to estimate the probability distribution of the region feature generated from the training images. Many other statistical learning approaches such as LDA [24], PLSA [15] and SVM [16] have also been used to improve image annotation.

In most generative model based approaches, the correlations between keywords are ignored to simplify the model calculations. Recently, it is realized that the correlations between annotated keywords can be used to improve the performance of image annotation. For instance, the keyword set \{sky, grass\} has a larger probability to be an image caption than \{ocean, grass\}. Only a few work has been done to investigate the word correlations, such as CLM [9] and WordNet-based approaches [10, 13]. The former employs the co-occurrence of words indirectly by using EM algorithm to fit a language model for generating the keywords set, while the latter makes use of the WordNet to exploit...
the hierarchy of the keywords. Recently, AGAnn [26] algorithm is proposed which incorporates WordNet and keyword correlations into a graph model for image annotation.

In this paper, we present a generative model based annotation method to label images with a subset of the vocabulary of the keywords. Obviously, the number of different sets of keywords is exponential with respect to the size of the vocabulary. To reduce the computation load, we propose an iterative greedy algorithm, which works as follows. First, in each iteration, we estimate the conditional probability of each keyword being the caption of the designated image. Our method computes the probability estimation using not only image feature, but also the correlation between the keyword and annotated keyword subset assigned in the previous iterations. After the probability estimation is finished, the keyword which brings the maximum annotation probability gain is selected to be added into the annotation subset. The “Automatic Local Analysis” is borrowed from the text information retrieval to estimate the correlations between keywords on the training set. A good feature generation estimation is very important to improve the annotation probability estimation in each iteration. Specifically, it will provide a good starting point for the initial iteration in our heuristic greedy algorithm. To improve the estimation of the image feature generation probability, we also present a new method for exploring the region similarity matching structure based on maximum weight bipartite graph matching algorithm.


Our contributions are as follows:

1. We propose a new heuristic greedy iterative image annotation algorithm based on the generative language model, which improves the annotation performance by incorporating the keyword correlations and the image region matching approach.

2. We exploit the “Automatic Local Analysis” of text information retrieval to analyze the correlations between keywords on the training set and incorporate it into our annotation algorithm.

3. We propose a new image generation probability estimation approach based on the maximum region similarity matching and statistical smoothing.

The rest of the paper is organized as follows. Section 2 introduces the related work in image annotation and the backgrounds of our work. Section 3 presents our greedy iterative image annotation framework, including the semantic correlation measure between keywords and the new region matching based region generating probability estimation. Section 4 introduces our heuristic annotation algorithm. We discuss the experiment results in Section 5. Section 6 concludes this paper.

2. RELATED WORK AND BACKGROUNDS

2.1 Related Work

A significant amount of research has been done to address the problem of image auto-annotation [2, 3, 4, 5, 9, 10, 12, 13, 14, 15, 16, 17, 21, 23, 24, 26, 27]. Both the generative model and the discriminant methods of machine learning have been applied to improve the annotation performance. Duygulu et al.[1] view the image annotation as a process of translating “visual language” to text. They utilize a machine translation model to collect the links between the words and image visual features, and use these links to annotate new image. Jeon et al. propose the CMRM [3] relevance model to estimate the joint probability of the image and the keywords. Latent semantic analysis (LSA) [14] and probabilistic latent semantic analysis (PLSA) [15] introduce latent variables to link image features with keywords. Florent et al. [15] build a linked pair of PLSA models to attach more importance to textual features. By viewing each annotation word as an independent class, and creating a different classification model for each word, text classification technique is applied to image annotation task [9]. Lately, Gao et al. [16] introduce a multi-resolution grid-based annotation framework for image content representation and a hierarchical boosting algorithm to address the problem in image annotation using classification technique. A fully automatic and high speed annotation system—ALIPR has been constructed by Li and Wang [17].

Some previous work demonstrated that keyword correlations can be utilized to improve the performance of image annotation. That is to say, we can use the subset of the annotation vocabulary to exploit the keyword correlations to get better image annotations. However, when the size of the vocabulary is large, it is computationally prohibitive to enumerate all the keyword subsets. Jin et al. [9] address the problem by using EM algorithm to fit a language model to generate an annotation keyword subset. However, the annotation speed is lower due to the EM algorithm. Munirathnam et al. [13] propose a hierarchical classification approach to image annotation. They use a hierarchy induced on the annotation words derived from WordNet [25]. Jin et al. [10] make use of the knowledge-based WordNet and multiple evidence combination to prune irrelevant keywords. Liu et al. [26] propose an adaptive graph model based on manifold ranking, which combines the WordNet and the pairwise keyword co-occurrence to improve the annotation effectiveness.

The measurement of the similarity between images or image regions is an important issue. In the discrete feature models, such as Translation Model [2] and CMRM [3], the image visual words vocabulary is generated by clustering and discretizing the region features. However, discretization leads to the loss of visual feature content. The continuous models, such as CRM [5] and MBRM [4], make use of the non-parameter kernel-based approach to improve the region generation probability estimation. However, these methods ignore the complexity of each image region contributing to the probability estimation of generating a certain image. The problem of region-based image similarity measure has been studied for several years in Content-Based Image Retrieval(CBIR). Various methods have been proposed to employ structural or mapping information to improve the retrieval performance [6, 7, 21, 22]. Wang et al. [6] define a weighted sum of region similarities to measure the similarity of images, the most similar regions have the top matching priority. In FUZZYCLUB [7], the similarity between a region and another designated image is valued by the smallest distance selected from all the distances computed from this
region to every other region of the designated image.

From the image content representation point of view, there are two widely accepted techniques in image annotation: region-based method or grid-based method. It is shown that employing global feature [23] or dividing the image into grid [4] can obtain good annotation result. However, this paper focuses on developing region segmentation and continuous feature estimation based image annotation, because of its potential value for practical application, such as annotation at the object level.

2.2 The Background of Relevance Model Based Image Annotation

Our annotation framework is based on the Relevance Model [3]. We will present a brief introduction to the Relevance Model based image annotation in this section.

For a given training set $T$, let $|T|$ denote the size of $T$. Each annotated image $I_i$ in the collection can be described using a set of image regions and annotation words: $J_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,m}; w_{i,1}, w_{i,2}, \ldots, w_{i,n}\}$, where $m$ and $n$ are the numbers of image regions and annotation words respectively, $f_{i,j}$ is the $D$ dimensional region feature vector, and $w_{i,j}$ is a binary variable, indicating whether or not the $j^{th}$ word appears in the $i^{th}$ image. Given a new image $I = \{f_1, f_2, \ldots, f_l\}$, the likelihood for a word $w$ to be an annotation keyword for $I$ is calculated as follows:

$$P(w|I) \propto P(w, I) = \sum_{i=1}^{|T|} P(w, I|J_i)P(J_i)$$

$$= \sum_{i=1}^{|T|} \prod_{j=1}^{|J_i|} P(f_{i,j}|I) \cdot P(w|J_i) \cdot P(J_i)$$

Then the best annotation is:

$$w^* = \arg \max_w P(w|I),$$

where $P(f_{i,j}|I)$ denotes the probability of region $f_j$ generated from training image $J_i$, and $P(w|J_i)$ denotes the probability of word $w$ generated from $J_i$. The assumption is that $P(J)$ is uniformly distributed.

3. THE ITERATIVE IMAGEANNOTATION FRAMEWORK

In this section, we first present the iterative function of our greedy algorithm, and then the keyword correlation analysis and the region matching based image generation estimation will be given.

3.1 The Iterative Function of the Annotation

We extend the Relevance Model based image annotation framework to keyword subset based annotation. Let $S_k$, $|S_k| = k$ denote a subset of the annotation vocabulary $V$. Given a new image $I$, the probability of $S_k$ to be annotated for $I$ is:

$$P(S_k|I) \propto P(S_k, I)$$

$$P(S_k, I) = P(w_k, S_{k-1}, I) = P(w_k|I, S_{k-1}) \cdot P(S_{k-1}, I)$$

where $S_k = S_{k-1} \cup \{w_k\}$, $k \geq 1$. Then the optimal annotation is:

$$S_k^* = \arg \max_{S_k \subseteq V} P(S_k, I).$$

Obviously, the number of different sets of keywords is exponential with respect to the size of the vocabulary. When $|V|$ is large, we can not solve it by enumerative search due to high computational cost. Therefore, we make use of a heuristic greedy algorithm to get an approximate optimal solution. According to Eqn.(3), we have the log-likelihood function as:

$$\log P(S_k, I) = \log P(S_{k-1}, I) + \log P(w_k|I, S_{k-1}).$$

Then we define the annotation function $f(S_k)$ as follows:

$$f(S_k) = \log P(S_k, I).$$

Because $f(S_k)$ and $P(S_k, I)$ have the same effect for ranking, the answer for $f(S_k)$ is the same as for $P(S_k, I)$. From Eqn.(5) and Eqn.(6), we have the iterative function as follows:

$$f(S_k) = f(S_{k-1}) + \log P(w_k|I, S_{k-1}).$$

Then we define the annotation probability gain of iteration $k$ as follows:

$$G_w(k) = f(S_k) - f(S_{k-1}) = \log P(w|I, S_{k-1}) = P(w|I, S_{k-1}).$$

The annotation gain $G_w(k)$ describes the ability of the word $w$ in boosting $P(S_k, I)$, where $P(S_k, I)$ is the probability of $S_k$ being the caption of image $I$. In order to archive an approximate optimal annotation, we use the following greedy heuristic: For the $k^{th}$ iteration, select keyword $w^*$ which maximizes $G_w(k)$ and add it into the $S_k$. Thus, the objective function of the $k^{th}$ iteration is:

$$w^* = \arg \max_{w \in V \setminus S_{k-1}} G_w(k).$$

For simplicity, we assume $I$ and $S_{k-1}$ are independent, then we have:

$$P(w|I, S_{k-1}) = \frac{P(w|I)P(w|S_{k-1})}{P(w)}.$$  

Assuming that $P(w)$ is uniformly distributed, we have:

$$w^*_k = \arg \max_w P(w|I)P(w|S_{k-1}) = \arg \max_w P(w|I)P(w|S_{k-1}).$$

Note that the maximum likelihood estimation for $P(w|S_{k-1})$ is:

$$P_M(w|S_{k-1}) = \frac{\#\{J|w, S_{k-1} \in J\}}{\#\{J|S_{k-1} \in J\}},$$

where $\#\{J|w, S_{k-1} \in J\}$ denotes the number of images in which the keyword $w$ and set $S_{k-1}$ appear together. For a limited training set, when $|S_k|$ is large, the co-occurrence of $w$ and $S_{k-1}$ is rare, which means the above probability has many zeros. However, a zero probability event in the training set does not mean it never happen in the future, thus smoothing is necessary.

In the text information retrieval, smoothing is usually performed by making use of a large background collection to assign a non-zero probability to the unhappened event in current model. For instance, we can choose a larger training image set for smoothing. However, according to Zipf’s law, there are many words appearing in a small subset of the images [13], which means the sparseness cannot be solved by
increasing the size of the training set. Due to the above reason, we introduce a modified smoothing method as follows:

We define \( \text{relation}(w, S_{k-1}) \) as the measurement of the correlation between keyword \( w \) and keyword set \( S_{k-1} \):

\[
\begin{align*}
\text{relation}(w) &= 1, |S_k| = 1 \\
\text{relation}(w, S_{k-1}) &= \sum_{w' \in S_{k-1}} \text{Sim}(w, w'), |S_k| > 1 ,
\end{align*}
\]

where the calculation of \( \text{Sim}(w, w') \) is described in the following subsection which reflects the semantic similarity between \( w \) and \( w' \). Note that the \( \text{relation}(w, S_{k-1}) \) should be normalized to satisfy the property of probability which is:

\[
P(w|S_{k-1}) = (1 - \gamma)P_M(w|S_{k-1}) + \gamma\text{relation}(w, S_{k-1}),
\]

where \( \gamma \) is the smoothing factor. When \( |S_k| \) is large, the \( \text{relation}(w, S_{k-1}) \) will dominate the probability \( P(w|S_{k-1}) \), so that \( \gamma \) will be close to 1.

The probability \( P(w, I) \) is computed as the expectation over the images in the training set, as in MBRM [4]. Since each word appears in an image only once, it would be more appropriate to describe annotation words with Bernoulli distribution:

\[
p(w, I) = \sum_{i=1}^{|T|} p(w, I|J_i) p(J_i) = \sum_{i=1}^{|T|} \Pi_{j=1}^m p(f_j|J_i) p(w|J_i) \Pi_{w \neq w'} (1 - p(w'|J_i)).
\]

Meanwhile, a beta prior (conjugate to a Bernoulli) is applied for smoothing.

3.2 The Words Semantic Correlation Measurement

The “Automatic Local Analysis” (ALA) [11] is an effective tool to find the correlations between keywords (terms) by using the term co-occurrence matrix. If we regard each training image as a document containing the labeling keywords, a similar approach of ALA can be applied to measure the semantic similarities between image captions. The approach is described as follows. First, each training image is represented with its annotated keywords. Then, the relationship between the captions and the images is described as the association matrix \( M_{n \times |T|} \), where the \( i \)th row denotes the pattern of word \( w_i \) occurring in the training set, the \( j \)th column denotes the annotation of the image \( J_j \), \( M_{ij} \) denotes whether the \( i \)th word occurs in the annotation of the \( j \)th image.

Generally, the matrix \( M \) will be very sparse. Therefore, we first take the entire training set as the background collection to smooth the probability with Jelinek-Mercer algorithm [15]. It allocates a small nonzero probability to the keywords whose corresponding occurring probability is zero. Then, we make use of the co-occurrence pattern to measure the words semantic correlations. The semantic correlation between \( w_u \) and \( w_v \) is defined as follows:

\[
c_{w_u, w_v} = \sum_{J_j \in T} M_{uj} \times M_{vj}
\]

For normalization, we have:

\[
s_{w_u, w_v} = \frac{c_{w_u, w_v}}{c_{w_u, w_u} + c_{w_v, w_u} - c_{w_u, w_v}},
\]

where \( s_{w_u, w_v} \) measures the co-occurrence frequency of \( w_u \) and \( w_v \).

According to Eqn.(16), the “neighborhood” of \( w_u \) can be defined as: \( \tilde{s}_{w_u} = (s_{w_u, w_1}, s_{w_u, w_2}, ..., s_{w_u, w_n}) \), then we can judge the correlations of \( w_u \) and \( w_v \) by their “neighborhood”. Specifically, if \( w_u \) and \( w_v \) have similar “neighborhood” (they appear in similar context), they are more likely to be semantic correlated. The cosine of \( \tilde{s}_{w_u} \) and \( \tilde{s}_{w_v} \) is applied to measure the similarity:

\[
\text{Sim}(w_u, w_v) = \frac{\tilde{s}_{w_u} \cdot \tilde{s}_{w_v}}{|\tilde{s}_{w_u}| \times |\tilde{s}_{w_v}|}
\]

For instance, there are five words in the annotation vocabulary \( V = \{ \text{sky, ocean, grass, tiger, horse} \} \), four images in the training set, \( T = \{ J_1, J_2, J_3, J_4 \} \). The association matrix \( M \) is defined in Fig.1. Each row in the matrix can be regarded as a vector \( \vec{s} \), the semantic correlation between words “grass” and “sky” can be measured by Eqn.(17). Although “tiger” and “horse” seldom co-occur together, they still have high correlation, since they both belong to “animal”.

3.3 Feature Estimation based on Maximum Weight Region Matching

From the image generation point of view, the probability of image \( I \) generated from image \( J \) is determined by the joint probability of each region in \( I \). Therefore, the image region generation estimation is a fundamental issue for image auto-annotation.

In CRM [5] and MBRM [4], the probability of region \( f_j \) generated from image \( J \) can be measured as the average distance(similarity) between region \( f_j \) and each region \( g_k \) in the image \( J \). On the one hand, the measurement ignores the semantics of the salient region after segmentation. For example, see Fig.1 (a), the similarity between the “tiger” in the two images should be determined by the corresponding “tiger - tiger” comparison, rather than the “tiger - background” comparison. On the other hand, the image segmentation algorithms often produce a considerable amount of regions which are meaningless or with ambiguous meanings. For example, the marked regions in Fig.2(b) are some constituent parts of the object “house” and “beach”. Due to their weak semantics, they are readily to obtain higher visual similarity with other regions. Consequently, the gen-

Figure 1: The words association matrix M

<table>
<thead>
<tr>
<th></th>
<th>J1</th>
<th>J2</th>
<th>J3</th>
<th>J4</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1(sky)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>w2(ocean)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>w3(grass)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>w4(tiger)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>w5(horse)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
estimation value of region $b_i$ generated from $b'_j$:

$$w_{i,j} = \frac{\exp\left[-(g_i - f_j)^T \Sigma^{-1} (g_i - f_j)\right]}{Z_w},$$  \hspace{1cm} (18)$$

where $g_i, f_j$ are visual feature vectors of $b_i, b'_j$. $Z_w$ is the normalization factor.

In the following discussions, we introduce the definitions of maximum weight matching of complete bipartite graph. A matching of the complete bipartite graph $G$ denoted as $M = \{i, \pi(i)\}, N_i \in N, N'_{\pi(i)} \in N'$. $M$ is a subset of $E$ such that no two edges of $M$ share a common vertex. A maximum weight matching of the complete bipartite graph $G$ is a matching where each node of $N$ is in this matching and the sum of the weights of the edges of the matching is maximized, that is:

$$\max \left(\sum_{i=1}^{m} w_{i, \pi(i)}\right).$$  \hspace{1cm} (19)$$

The above definition assumes that the number of vertexes in $N$ and $N'$ are the same. However in image annotation, the number of regions in different images are not always the same. When the number of regions in image $I$ is larger than that in image $J$, the maximum weight matching cannot be found according to the above definition. Therefore, we modify the condition as: all nodes share the least amount of common nodes instead of no two edges sharing a common vertex.

The maximum weight bipartite graph matching could be solved by Hungarian algorithm [8], which can also handle the situation that the numbers of nodes in set $N$ and $N'$ are not the same. The running time of the algorithm is determined by the number of nodes. Generally, the number of regions segmented from a image is very limited. For instance, in the ECCV2002 benchmark, the number of regions of the image is at most 10. Thus the bipartite graph matching algorithm can run efficiently. The probability of $b_i$ generated from $J$ is determined by the non-parameter estimation value of $b_i$ with its optimal matching $b_{\pi(i)}$ found by Hungarian algorithm, which is given by:

$$P_{opt}(b_i|J) = w_{i,k}, \hspace{0.5cm} k = \pi(i).$$  \hspace{1cm} (20)$$

According to the above Equation, the probability is estimated by only one sample – its optimal matching region. The sample set is sparse which will lead to low bias but high variance [18]. We make use of Jelinek-Mercer method for smoothing. Specifically, all the regions of the image are selected as background collection to calculate the probability $P_b(f_j|I_i)$, then we have the smoothed $P(b_i|J)$ as follows:

$$P(b_i|J) = \lambda P_{opt}(b_i|J) + (1 - \lambda) P_b(b_i|J),$$  \hspace{1cm} (21)$$

where $\lambda$ is the smoothing coefficient, its best value could be determined by the validation set.

4. THE ANNOTATION ALGORITHM

In this section, we will present our iterative annotation algorithm incorporating the words correlation measurement and the maximum weight matching based region generating probability estimation. The annotation algorithm is:

According to Eqn.(14), the computation of $P(w|S_{i-1})$ is dominated by relation($w, S_{i-1}$). That is to say, in the first iteration, the selection of the first annotation keyword only
depends on the estimation of the joint probability of the word \( w \) and the image \( I \). While Eqn.(15) implies that the joint probability is closely related with the region feature generating probability \( P(f_j|J_i) \), which means improving the accuracy of the estimation of \( P(f_j|J_i) \) will bring a good starting point of the annotation algorithm. In the subsequent iterations, the correlations between keywords will be combined into the annotation process to improve the performance.

5. EXPERIMENTS
The merit of our method is that the correlations between keywords and the region matching based image feature generation estimation are both readily combined into the image annotation by a heuristic greedy algorithm, and both of their contributions to the image annotation can be extensively exerted. In order to demonstrate the effectiveness of the proposed approach, we design the following 3 experiments:

- The first experiment tests the overall effectiveness of our iterative annotation algorithm. Our annotation approach is mainly compared with MBRM [4], due to its better performance and representation.

- The second experiment investigates the effectiveness of the proposed semantic correlation measurement in improving the image annotation. In this experiment, the feature generation estimation is same as MBRM [4].

- The third experiment evaluates the effectiveness of the proposed feature generation probability estimation using the maximum weight matching and smoothing technique.

5.1 The Experiments Setup
We test our algorithm using Corel data set provided in [2], which consists of 5000 images from 50 Corel Stock Photo CDs, each CD includes 100 images on the same topic, and each image is linked with 1-5 annotation words. There are totally 373 different words in the data set. We divide the data set into three parts: the training set which consists of 4000 images, the validation set with 500 images from 50 CDs and the test set with the remaining 500 images. The validation set helps to determine the model parameter. After the parameter is decided, we combine the validation set with training set to form a new large training set and retrain the model. We compute the recall and precise to measure the performance of the algorithm. Given a query word \( w \), let \( |W_G| \) denote the number of human annotated images with label \( w \) in the test set, \( |W_M| \) denote the number of annotated images with the same label by our algorithm, and \( |W_C| \) denote the correct annotations by our algorithm. The recall and precise are defined as: \( \text{Recall} = \frac{|W_G|}{|W_M|}, \text{Precision} = \frac{|W_G|}{|W_C|}. \) Recall measures the completeness of the images with annotation \( w \); Precision measures the accuracy of the images with annotation \( w \). The average recall and precision over all the words evaluate the system performance.

5.2 Experimental Results
We compare our work mainly with MBRM, both of the approaches are based on image segmentation and the Gaussian kernel density feature estimation. Our approach and the compared approach use the same training set and the test set; the size of each annotation set is fixed at 5; the average recall and precision are calculated over all 263 words which appear in the test set.

1. The overall performance of our algorithm
Fig.4 compares the Anno_iter approach proposed in this paper and MBRM. According to the figure, the average recall and precision both improve significantly which are 21% and 11% better than MBRM respectively. Specifically, the recall and precision increase from 16.1% and 19.0% to 19.5% and 21.7%, respectively.

2. The effectiveness of word correlations measure
In this experiment, the probability estimation method is the same as MBRM. However we apply word correlation measure to our annotation framework. Fig.5 shows the experimental results for MBRM and our algorithm only using Words Correlation(WCor). According to the figure, the average recall and precision increase from 16.1% to 18.6% and from 19.0% to 19.7% respectively, which demonstrates our words semantic correlation measure could be an effective way to smooth the maximum likelihood estimation to exploit the words correlation in the annotation process.

3. The Effectiveness of Maximum Weight Matching based Probability Estimation

![Algorithm 1 Anno_iter](image-url)
We compare the annotation performance of our algorithm using Maximum Weight Region Matching (MWRM) with two other algorithms: the discrete feature model CMRM, and the continuous feature model MBRM. To make sure that the word correlation does not affect the result, we keep the words independent in the MWRM-based annotation. The smoothing factor $\lambda$ is set to 0.7 for best performance. The result is shown in Fig. 6. Apparently, the continuous feature estimation outperforms the discrete feature model, and the performance of MWRM is much higher than the initial non-parameter estimation MBRM. The reason is that the CMRM loses too much image content by using feature discretization, while our method MWRM successfully reduces the influence of the trivial regions through optimal region matching structure, thus improves the annotation performance. Specifically, compared with the MBRM, the recall increases from 16.1% to 18.3%, and the precision increases from 19.0% to 19.8%.

4. Annotation Examples

Fig. 7 lists some annotation examples annotated by our algorithm. The word correlations play an important role in choosing the annotation of the first three images in this Figure. Take the first image as an example, “forest” obviously has higher correlation with “tiger” than “hut” and “water” do, which helps our annotation algorithm find the correct annotation “forest”. It is verified that considering word correlations can help find high-related words in the annotation process and can lead to higher annotation accuracy.

Consider the last two images, there are almost no correct words appeared in the annotation result of MBRM, while almost all the ground-truth words appear in our annotation result. The reason is that our MWRM method can compute the region generation probability more accurately, which results in more accurate probability estimation at the image level. In addition, our iterative algorithm can incorporate the word correlations to help find the most relevant words based on the MWRM generation probability estimation to improve the image annotation significantly.

From the experiment result, it can be concluded that: the word correlations and MWRM-based probability estimation can improve the image annotation quality; our annotation algorithm can incorporate both of them to improve the image annotation more significantly.

6. CONCLUSIONS

In this paper, we propose a novel annotation algorithm which extends the current annotation framework to word set based annotation. To avoid the difficulty of computation, we propose a heuristic greedy algorithm to implement the annotation by investigating the semantic similarity between keywords. Because of its fundamental role, we also introduce a new approach to estimate the region feature probability distribution by maximum weight bi-graph matching algorithm. The experiment results demonstrate that our annotation algorithm can combine both the word correlations and the region matching approaches to improve the performance of the image annotation significantly.

In the future work, we plan to make use of more word correlations and hierarchies learned from training set and other knowledge resource, such as WordNet, and incorporate them into our annotation algorithm to improve the annotation performance further. In addition, we will extend our work to image annotation and retrieval at the semantic object level.

7. REFERENCES

Figure 7: Annotation examples