Exploring Flickr’s Related Tags for Semantic Annotation of Web Images

Hongtao Xu
Fudan University
Shanghai, China
061021054@fudan.edu.cn

Xiangdong Zhou
Fudan University
Shanghai, China
xdzhou@fudan.edu.cn

Mei Wang
National University
Singapore
wangmei@comp.nus.edu.sg

Yu Xiang
Fudan University
Shanghai, China
072021109@fudan.edu.cn

Baile Shi
Fudan University
Shanghai, China
bshi@fudan.edu.cn

ABSTRACT

Exploring social media resources, such as Flickr and Wikipedia to mitigate the difficulty of semantic gap has attracted much attention from both academia and industry. In this paper, we first propose a novel approach to derive semantic correlation matrix from Flickr’s related tags resource. We then develop a novel conditional random field model for Web image annotation, which integrates the keyword correlations derived from Flickr, and the textual and visual features of Web images into an unified graph model to improve the annotation performance. The experimental results on real Web image data set demonstrate the effectiveness of the proposed keyword correlation matrix and the Web image annotation approach.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]

General Terms

Algorithms, Measurement, Experimentation

Keywords

Web image annotation; Flickr’s tag; Keyword correlation; Conditional random field model

1. INTRODUCTION

The mismatch between low level visual features and high level semantics, the so-called *Semantic Gap* problem [13], has posed great challenges to the content based multimedia applications. Recent research efforts have suggested that the relationship between semantics is one of the important clues to mitigate the difficulty of this problem.

Among the variants of content based multimedia applications, exploiting the semantic correlations brings promising improvement to the performance of Automatic Image Annotation (AIA). For instance, the semantic correlation provides strong hints that the keyword set {*sky, grass*} has a larger probability to be an image label than {*ocean, grass*}. Some previous research efforts estimate the semantic correlations between keywords according to the frequency of appearance of keywords in training set or some lexicons, such as WordNet [3]. However, the use of limited training data [25, 30] or WordNet [10, 21] is often ineffective for problems with unconstrained vocabulary such as the Web image collection. This is because many visually co-occurring terms in the Web collection may not appear in the training set or WordNet.

With the rapid development of Web social community, the applications which exploit the social media resources, such as Flickr [1] and Wikipedia, have become popular and attracted much attention from both academia and industry [20]. Many recent research efforts explore correlations between keywords derived from these resources to infer image semantics. Wu et al. [27] proposed a new Flickr distance to measure the visual similarity between concepts according to Flickr. Schmitz [19] proposed the building of facted ontology from Flickr’ tagging resources. Wang and Domeniconi [26] proposed deriving semantic kernel from Wikipedia for text classification. Differing from previous works which crawl a huge amount of Web pages for parsing and analysis [27, 26], we propose to directly explore the Flickr’s Related Tags (RT)[1] resources. That is we view Flickr as a Web-Scale Image Semantic Space, and submit the keywords of annotation vocabulary to Flickr to obtain the returned Related Tag (RT) set to construct RT Graph in pure keyword space. We then derive a keyword correlation matrix from the RT graph to explore the relationship between semantics.

However, merely relying on keyword correlations to infer image semantics is often limited. This is because it covers only one aspect of image features and ignores the structures of associated texts and visual features of images. In fact, the associated texts of Web images, which include image file name, ALT texts, captions, surrounding texts and page title,

CIVR ’09, July 8-10, 2009 Santorini, GR
Copyright 2009 ACM 978-1-60558-480-5/09/07 ...$5.00.

[1] RT can be obtained by using Flickr’s APIs: flickr.tags.getRelated. It returns “a list of tags ‘related’ to the given tag, based on clustered usage analysis ”–refer to: http://www.flickr.net/services/api/flickr.tags.getRelated.html

Table 1: Examples of Flickr's Related Tags

Tag	Flickr's Related Tags
flower	yellow pink red nature spring green insect purple plant garden rose white bee orange closeup tulip blue sunflower color water lily petals daisy summer flor leaf pollen tree
beach	sea sand ocean water sunset sky sun wave summer cloud blue landscape rock surf girl coast vacation reflection boat california island seaside pacific holiday shore people wave travel woman dog pier light
car	road street auto classic night red automobile city reflection sky beetle driving blue traffic rain urban trees highway water people motion building cloud vehicle sunset trip

etc. cover the contents of the corresponding Web image to a certain degree. Therefore, the use of visual features together with different types of associated texts should provide valuable information to semantic inference of Web images. However, it has been demonstrated that the inference of different kinds of associated texts on the semantics of Web images varied greatly depending on the class of images. Moreover, the interactions among these associated texts also play an important role [28]. Although many previous works have utilized information on associated texts in their research, they either assign fixed weights to different types of associated texts heuristically [14, 18], or view textual and visual features as orthogonal sources [5, 24] by using traditional AIA models for semantic inference.

In order to integrate the vast array of information available in correlated keywords, associated texts and visual features of Web images for Web image annotation, we propose a conditional random fields (CRF) model to adaptively and systematically model all available information in an unified framework. More specifically, we define various types of cliques and the corresponding potential functions to represent the semantic contributions of different types of features in image semantic inference, and integrate them into the CRF model uniformly.

Our contributions are as follows:

1. We exploit the popular Web Photo Community site Flickr as a Web-Scale Image Semantic Space to analyze the semantic correlations between keywords and incorporate it into our annotation framework.

2. We propose a conditional random field model based Web image semantic annotation framework to adaptively and systematically integrate various information sources of Web images.

We conduct experiments on a real Web image data set to demonstrate the effectiveness of our proposed CRF based Web image annotation approach and the Flickr based keyword correlation measurement method.

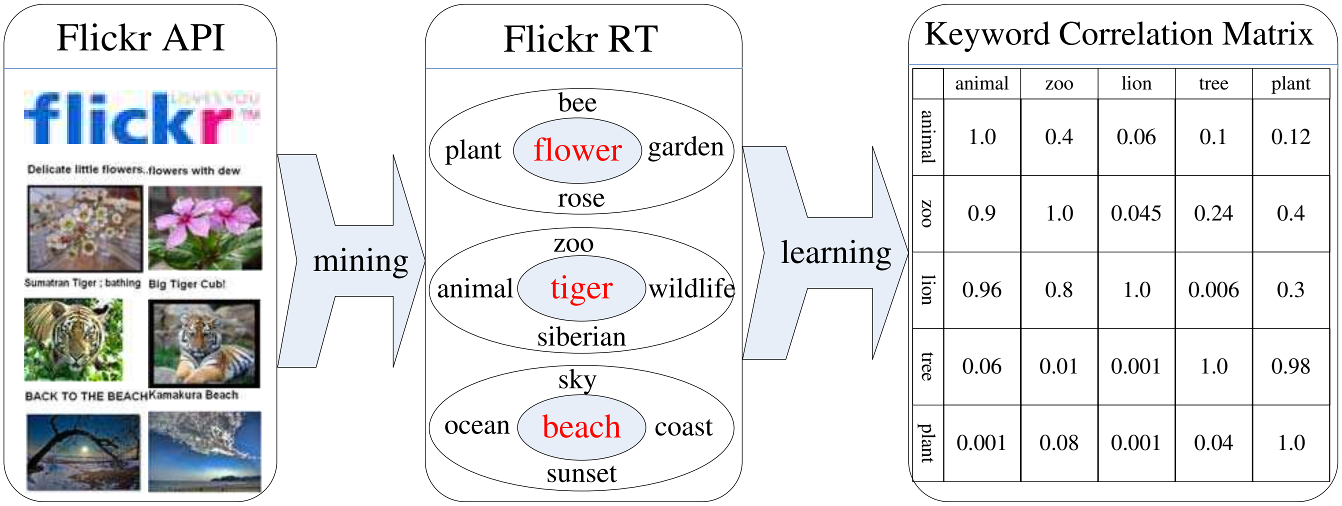
The rest of this paper is organized as follows. Section 2 gives related work. Section 3 introduces keyword correlation matrix derived from Flickr. Section 4 presents the CRF based Web image annotation algorithm. We discuss the experimental results in Section 5. Section 6 concludes this paper.

2. RELATED WORK

In recent years, much attention has been paid to the research on AIA. Various of machine learning techniques or statistical models have been employed to develop a variety of AIA models. Essentially, the AIA models can be divided into two main categories, namely the probabilistic model based methods and classification based methods. The first category focuses on inferring the correlations or joint probabilities between images and keywords. The representative work include Translation Model(TM) [4], CMRM [8], CRM [12], MBRM [7], multiple segmentations based AIA [23] etc. The classification based methods try to associate keywords or concepts with images by learning classifiers. Methods like SVM-based approaches [2], multi-instance learning [29] fall into this category. However, the existing approaches do not focus on annotating Web images and often neglect the available textual information of Web images.

Sanderson and Dunlop [18] were among the first to model image contents using a combination of texts from associated Web pages. Li et al. [14] proposed a search and mining framework to tackle the AIA problem. Given an unlabeled image, content-based image retrieval (CBIR) was firstly performed to find a set of visually similar images from a large-scale image database. Then clustering was performed to find the most representative keywords from the annotations of the retrieved image subset. These keywords, after saliency ranking, were used to annotate the unlabeled image. Feng et al. [5] described a bootstrapping framework by adopting a co-training approach involving classifiers based on two orthogonal set of features–visual and textual. Tseng et al. [24] built two models based on image visual and textual features, and weighted them to annotate the unlabeled Web images. Xu et al. [28] presented a method to adaptively model the distributions of the semantic annotation keywords on the associated texts of the Web image.

Some previous research efforts demonstrated that keyword correlations can be utilized to improve the performance of image annotation. Jin et al. [9] addressed the problem by using EM algorithm to fit a language model to generate an annotation keyword subset. Srikanth et al. [21] proposed a hierarchical classification approach for image annotation. They used a hierarchy induced on the annotation keywords derived from WordNet. Jin et al. [10] made use of the knowledge-based WordNet and multiple evidence combination to prune irrelevant keywords. Zhou et al. [30] proposed an iterative image annotation approach by exploring keyword correlations. Tang et al. [22] proposed a graph-based learning approach SSMR to measure the pairwise concept similarity. However, these approaches usually learn the keyword correlations according to the occurrences of keywords in the training set or lexicon, and the correlation may not reflect the real correlation for annotating Web images. With the rapid development of Web social community, many applications have emerged that exploit the social media resources, such as Flickr and Wikipedia [20]. Wu et al. [27] proposed a new Flickr distance to measure the visual similarity between concepts according to Flickr. Flickr distance aims to describe the concepts' semantic distance in the visual sense. However, Flickr's tags correlation in the keyword space is ignored. Schmitz [19] proposed building ontology from Flickr' tagging resources. Wang and Domeniconi [26] proposed deriving semantic kernel from Wikipedia for text classification. Compared with Flickr, it seems that



	animal	zoo	lion	tree	plant
animal	1.0	0.4	0.06	0.1	0.12
zoo	0.9	1.0	0.045	0.24	0.4
lion	0.96	0.8	1.0	0.006	0.3
tree	0.06	0.01	0.001	1.0	0.98
plant	0.001	0.08	0.001	0.04	1.0

Figure 1: The process of deriving the keyword correlation matrix from Flickr's Related Tag (RT).

Wikipedia lacks of the description of co-occurrence of concepts in visual sense. In this paper, we propose to learn the keyword correlation matrix by exploiting Web social media, such as Flickr's Related Tag resource.

Due to its powerful modeling ability, graph model has been applied in AIA. For instance, Liu et al. [15] proposed a graph model for adaptive image annotation. Feng and Manmatha [6] proposed a new kind of discrete visual feature and Conditional Random Field (CRF) model for image retrieval. It is evident that compared with traditional AIA, there are many types of associated information available for Web images. Therefore, differing from the previous work, we propose using a unified graph model to adaptively and systematically model all available information of Web images for the semantic annotation.

3. KEYWORD CORRELATION MATRIX DERIVED FROM FLICKR

Flickr [1] is a popular Web photo community site, which enables users to manage and share digital photos. Flickr tags are user-generated labels for images. According to Flikcr's Related Tag API, each tag has a list of "related" tags, obtained by usage analysis. Table 1 shows some keywords and their Related Tags samples by using Flikr's Related Tag API. It is apparent that we can obtain the keyword (tag) correlations matrix or some lexicon by mining the Related Tags, which has been seen very valuable for a broad applications [26]. Figure 1 gives the process of deriving keyword correlation matrix from Flickr.

3.1 Flickr's Related Tags Graph

We view Flickr as a Web-scale Image Semantic Space (WISS). To build a local keyword semantic subspace for our annotation task, we submit the annotation vocabulary to Filckr to retrieve a keyword subset, which is composed of the returned Related Tags (RT)(or concepts). When we get these RT resource, the first important issue is to deal with the verbosity and spamming problem which is prone in Web resource. On close inspection, we found that Web tagging resources have many noisy tags (concepts), such as "a123". These noisy tags do not have explicit semantics, and should be removed. In this work, in order to alleviate the verbosity and spamming problem, we employ stop word pruning and noun word extraction tools from WordNet [3] to filter the Related Tags obtained from Flickr.

To better model the keyword correlations contained in the Flickr's related tags, we build a directed Related Tags graph $G_{RT} =< V', E' >$. Here the vertex set V' consists of keywords in the Image Semantic Space, and a directed edge from w to w' is denoted by $e_{ww'} \in E'$ which is established if and only if $w' \in RT(w)$, where $RT(w)$ is the set of Related Tags (RT) of w.

3.2 Topic based Keyword Correlation Matrix

The Graph G_{RT} is very sparse, which gives rise to difficulty in generating the keyword semantic correlation matrix. Obviously keywords characterizing the contents of similar images often belong to the same topic. Thus one way to alleviate the sparsity problem is to use topics as the basis to model images' contents. To this end, we employ Fisher Discriminant Analysis (FDA) to analyze the keywords grouped by common topic. For example, "*beach, ocean, coast, sea, ...*" often characterize the images about "beach", and "*tiger, animal, lion, zoo, ...*" often characterize the images about "zoo", so they can be categorized into two different semantic topics, which can be used to improve the keyword correlation estimation.

The ***Keyword Correlation Matrix (KCM)*** is defined as follow: KCM is a $p \times p$ matrix, where p is the number of keywords, and $KCM(i, j) \in [0, 1]$ denotes the semantic correlation between the i^{th} and j^{th} keyword. In this Subsection we will discuss how to measure the keyword semantic correlation to generate the keyword correlation matrix.

Based on topic information and the definition of the Related Tags graph G_{RT}, we obey the following rules to extract keyword correlations:

- The keywords in the same topic are more similar than those belonging to different topics.
- For keywords w, w', the shorter the path from w to w' in $GraphG_{RT}$, the more similar between w and w'.
- The semantic correlations of the keywords which have

higher outside degrees in Graph G_{RT} should be penalized.

The semantic correlation between keyword w and w' is defined as follows:

$$KCM_T(w,w') = D_{topic}(w,w') * e^{-D_{G_{RT}}(w,w') \times \frac{degree_+(w)}{p}}, \quad (1)$$

where $D_{G_{RT}}(w,w')$ denotes the length of the shortest path from w to w' in graph G_{RT}, and $degree_+(w)$ is the outside degree of vertex w in G_{RT}. $D_{topic}(w,w')$ is the topic distance between the keywords w and w', which measures the contribution of topic information as follows:

$$D_{topic}(w_1,w_2) = \frac{\overrightarrow{v_{w_1}} \cdot \overrightarrow{v_{w_2}}}{\parallel \overrightarrow{v_{w_1}} \parallel \times \parallel \overrightarrow{v_{w_2}} \parallel}, \quad (2)$$

where $\overrightarrow{v_{w_1}}$ is the topic vector that contains the keyword w_1. Note that if w_1 and w_2 belong to the same topic, then the topic distance $D_{topic}(w_1,w_2)$ is maximized, which means that w_1 and w_2 is more similar than those keywords belonging to different topics.

3.3 Smoothing

Although Flickr provides more suitable information for keyword semantic correlation matrix estimation, it shows apparent bias for certain testing data set. On the other hand, the training image data set also contains valuable information for deriving keywords semantic correlation. Therefore, we combine the keyword semantic correlation matrix derived from Flickr and the one that derived from training data [30]. The combined matrix is defined as:

$$KCM = \lambda KCM_T + (1-\lambda)KCM_t, \quad (3)$$

where λ is a smoothing parameter, and KCM_t is the keyword correlation matrix derived from the training set.

4. CONDITIONAL RANDOM FIELD BASED WEB IMAGE ANNOTATION

Conditional Random Field (CRF) models [11] are undirected graphical models, which aim to provide a compact and flexible way to represent conditional model $P(X|Y)$, where both X and Y have non-trivial structure (often sequential). In this paper, we use CRF to model the generative distribution $P(w|I)$, that is the probability of the keyword w being the annotation of image I given the observation values of the features of image. We model the conditional probability using this formalism because it is illustrative and provides a unified framework to incorporate various features of Web images.

4.1 CRF based Annotation Object Function

For a given training image set L_{train}, each labeled image $J \in L_{train}$ can be represented by $J = \{W, V, T\}$, where the annotation keywords W is a binary annotation keyword vector indicating whether a keyword is the annotation of image J; $V = \{f_1, \ldots, f_m\}$ is a set of region-based visual features of image J; and $T = \{T_1, \ldots, T_n\}$ is a set of the textual features of image J.

The conditional random field is constructed from an undirected graph G. As shown in Figure 2, the vertex set in G consists of three types of nodes: the keyword node An_i, the textual feature nodes T_i and the visual feature nodes f_i. The edges define the semantic relationships between the nodes.

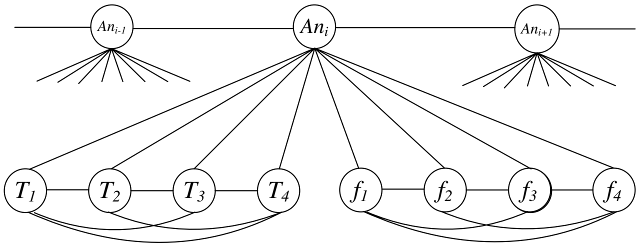


Figure 2: The proposed Conditional Random Field model, where T_i is the textual feature, f_i is the region-based visual feature, and $An_i = An_{i-1} \cup w_i$.

The probability of keyword w being the annotation of image I can be estimated as follows:

$$P(w|I, An_i) = \frac{1}{Z_\beta} e^{-\sum_{c \in C} \psi_c(c;\beta)}, \quad (4)$$

where An_i is the known annotation keywords of I, C is the cliques set, each $\psi(\cdot,\beta)$ is a non-negative *potential function* over clique configurations parameterized by β, and Z_β normalizing the distribution:

$$Z_\beta = \sum_{w,I} e^{-\sum_{c \in C} \psi_c(c;\beta)}. \quad (5)$$

Then the i^{th} best annotation keyword is:

$$w_i^* = argmax_w P(w|I, An_{i-1}). \quad (6)$$

Eqn.4 shows that the probability distribution is uniquely defined by the cliques set C and the potential function ψ.

4.2 Clique Set Type

In this work, we focus on the following five types of clique sets:

- ***Singleton keyword term(SKT)***: the clique set containing the singleton keyword node, which acts as a form of annotation keyword prior.
- ***Single textual term(STT)***: the clique set containing the keyword node and exactly one textual node.
- ***Multiple textual terms(MTT)***: the clique set containing the keyword node and two or more textual nodes.
- ***Single visual term(SVT)***: the clique set containing the keyword node and exactly one visual node.
- ***Multiple visual terms(MVT)***: the clique set containing the keyword node and two or more visual nodes.

4.3 Potential Function

Potential function ψ plays a very important role in the probability estimation. Here we focus on five types of potential functions corresponding to the five types of clique sets. We follow the common convention and parameterize the potentials as follows:

$$\psi_c(c;\beta) = \lambda_c f(X_c, Y_c), \quad (7)$$

where f is a real-valued *feature function* over cliques and λ_c is the weight given to the particular feature function. In the remainder of this subsection we will specify the *potential* *functions* we used, and we will show how to automatically determine the weights in the following subsection.

The first type of clique set we used is the *singleton keyword term* (SKT) clique set. A *potential function* over such a clique should measure the probability of the keyword w as the annotation of Web image I on the condition of the prior An_i. So we define the potential function as follows:

$$\psi_{SKT}(c) = \lambda_c f_{SKT}(X_c, Y_c) = \lambda_c logCor(w, An_i), \qquad (8)$$

where $|An_i|$ is the size of annotation keyword set An_i, and $Cor(w, An_i)$ measures the correlation between the keyword w and annotation keyword set An_i. This correlation can be computed based on the keyword correlation matrix KCM derived from Flickr:

$$Cor(w, An_i) = \frac{1}{|An_i|} \sum_{w_i \in An_i} weight(w_i) \times KCM(w_i, w), \quad (9)$$

where $KCM(w_i, w)$ is the semantic correlation between w_i and w, which has been introduced in Section 2. $weight(w_i)$ denotes the weight of w_i, and $weight(w_1) = 1$, $weight(w_i) = \rho \times weight(w_{i-1}), i = 2, \ldots, k$, where $\rho \in [0, 1]$ is the weight shrinkage factor.

The *single textual term* (STT) clique is 2-clique consisting of an edge between a textual feature T_i and w. A *potential function* over such a clique should measure how likely T_i describes the semantic of w. We can define this type of potential function as:

$$\begin{aligned}\psi_{STT}(c) &= \lambda_c f_{STT}(X_c, Y_c) = \lambda_c logCor(w, T_c) \quad (10)\\ &\propto \lambda_c log\{(1-\alpha)\frac{tf_{w,T_c}}{|T_c|} + \alpha\frac{df_w}{|D|}\},\end{aligned}$$

where $tf_{w,T}$ is the number of times keyword w occurring in T, $|T|$ is the total number of keywords in T, df_w is the number of w occurring in all types of texts, and $|D|$ is the total number of keywords in all types of texts. α is the smoothing parameter. This potential function makes the assumption that the more likely a keyword fits the language model of text T, the more likely T describes the semantic of w.

Next, we consider *multiple textual term* (MTT) cliques that contain two or more textual nodes. For this purpose, we construct a *potential function* over cliques that consist of the set of two or more texts $\Sigma = \{T_i, \ldots, T_j\}$ and the keyword w. Such potential functions have the following form:

$$\begin{aligned}\psi_{MTT}(c) &= \lambda_c f_{MTT}(X_c, Y_c) = \lambda_c logCor(w, \Sigma) \quad (11)\\ &\propto \lambda_c log\{(1-\alpha)\frac{\sum_{T_l \in \Sigma} tf_{w,T_l}}{\sum_{T_l \in \Sigma}|T_l|} + \alpha\frac{df_w}{|D|}\}.\end{aligned}$$

The *potential function* of the *single visual term* (SVT) clique measures how likely the visual feature f_i describes the semantic of w. The potential function is defined as:

$$\begin{aligned}\psi_{SVT}(c) &= \lambda_c f_{MVT}(X_c, Y_c) = \lambda_c logCor(w, f_i) \quad (12)\\ &\propto \lambda_c log\{\sum_{i=1}^{K} P(f_i|J_i)P(w|J_i)P(J_i)\},\end{aligned}$$

where K is the number of images in the neighborhood of image I. $P(w|J_i)$ denotes the probability of keyword w being generated from J_i, which can be estimated by maximum likelihood estimation. Further, we assume that $P(J)$ is uniformly distributed. $P(f_i|J_i)$ is the probability of the image region f_i being generated from J_i [30].

Original image	**samples**		
Beach, ocean, sky	**Label:** Beach **Prior:** ocean,sky	**Label:** ocean **Prior:** beach, sky	**Label:** sky **Prior:** beach, ocean

Figure 3: Example of generating training samples.

Lastly, we define the *potential function* of *multiple visual term* (MVT) clique. A *potential function* over such a clique should measure how likely the regions set $B = \{f_i, \ldots, f_j\}$ describes the semantic of w. The potential function is defined as:

$$\begin{aligned}\psi_{MVT}(c) &= \lambda_c f_{MVT}(X_c, Y_c) = \lambda_c logCor(w, B) \quad (13)\\ &\propto \lambda_c log\{\sum_{i=1}^{K} \prod_{f_i \in B} P(f_i|J_i)P(w|J_i)P(J_i)\}.\end{aligned}$$

The number of cliques in MVT is exponential to the size of image regions. For simplicity, in this work, we only consider the cliques consisting of adjacent and sequential regions.

4.4 Parameter Estimation

Given our parameterized joint distribution and a set of potential functions, the final step is to set the parameter values β. Note that the clique set SKT consists of only one clique, so we have

$$\beta^T = \{\underbrace{\lambda_1}_{SKT}, \underbrace{\lambda_2, \ldots,}_{STT} \underbrace{\lambda_{i_1}, \ldots,}_{MTT} \underbrace{\lambda_{i_2}, \ldots,}_{SVT} \underbrace{\lambda_{i_3}, \ldots}_{MVT}\}. \qquad (14)$$

To estimate the parameter β, we first determine a neighborhood $neigh(I) \subseteq L_{train}$ of image I by using the generation probability estimation approach. In the generation probability estimation, we consider both the visual and textual features. Here we regard the pair of keyword and image as the training sample. As shown in Figure 3, if an image has k annotation keywords $\{w_1, \ldots, w_k\}(k > 1)$, we will build k training sample sets, and the $j^{th}(j <= k)$ sample set includes the annotation keyword w_j and the prior $\{w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_k\}$. For each sample image, we compute the value x_i of *feature function* corresponding to each clique, such as $x_1 = logCor(w, An_i)$. For convenience, we write the x_i in the vector form as:

$$x^T = \{\underbrace{x_1}_{SKT}, \underbrace{x_2, \ldots,}_{STT} \underbrace{x_{i_1}, \ldots,}_{MTT} \underbrace{x_{i_2}, \ldots,}_{SVT} \underbrace{x_{i_3}, \ldots}_{MVT}\}. \qquad (15)$$

Then $neigh(I)$ is represented as $L = \{y_i, x_i\}_{i=1}^{N}$, where x_i is a sample feature value, y_i is the observed label, and N is the number of the sample images in $neigh(I)$. The log-likelihood objective function is:

$$\begin{aligned}\mathcal{O}(\beta) &= \sum_{i=1}^{N} logP(y_i|x_i) \qquad (16)\\ &= \sum_{i=1}^{N} log\{\frac{1}{Z_\beta}e^{-\sum_{c\in C}\psi_c(c;\beta)}\}\\ &= -\sum_{i=1}^{N}\{\beta^T x_i + log\{\sum_{j=1}^{N} e^{-\beta^T x_j}\}\}\\ &= -\sum_{i=1}^{N}\beta^T x_i - Nlog\{\sum_{j=1}^{N} e^{-\beta^T x_j}\}.\end{aligned}$$

To maximize $\mathcal{O}(\beta)$, we set the derivatives to 0:

$$\frac{\partial\mathcal{O}(\beta)}{\partial\beta} = \sum_{i=1}^{N} x_i(-1 + N\frac{e^{-\beta^T x_i}}{\sum_{j=1}^{N} e^{-\beta^T x_j}}) = 0. \qquad (17)$$

To solve Eqn.17, we use Newton-Raphson algorithm, which requires the second-derivative or Hessian matrix:

$$\frac{\partial^2\mathcal{O}(\beta)}{\partial\beta\beta^T} = \qquad (18)$$

$$-N\sum_{i=1}^{N} x_i e^{-\beta^T x_i}\frac{x_i\sum_{j=1}^{N} e^{-\beta^T x_j} + \sum_{j=1}^{N} -x_j e^{-\beta^T x_j}}{(\sum_{j=1}^{N} e^{-\beta^T x_j})^2}.$$

Starting with β^{old}, a single Newton-Raphson update is:

$$\beta^{new} = \beta^{old} - (\frac{\partial^2\mathcal{O}(\beta)}{\partial\beta\beta^T})^{-1}\frac{\partial\mathcal{O}(\beta)}{\partial\beta}, \qquad (19)$$

where the derivatives are evaluated at β^{old}.

4.5 Auto-Generation of Training Set

As the basis of supervised Web AIA, we automatically generate training set using a heuristic method by mining the associated texts of Web images. The idea of generating the basic annotation is similar to the *term frequency* heuristic [17]. Here we consider two kinds of term frequency, that is, the frequency of keyword w appears in one type of the associated texts, and the frequency that accounts for the number of the associated texts types that w appears in. The basic idea is that keywords with higher frequency are more important to the semantic of the corresponding Web image. Here we denote the $i^{th}(i = 1, ..., m)$ type of associated texts as T_i. After filtering the stop words, the keyword set of the associated texts of image I is denoted as WS_I. For each keyword $w \in WS_I$, the confidence of w being the semantic annotation of image I is defined as follows:

$$Conf(w, I) = \frac{df(w)}{m} \times \sum_{i=1}^{m} \alpha_i * \frac{tf(w, T_i)}{|T_i|}, \qquad (20)$$

where $df(w)$ refers to the number of T_i that w appears in; $tf(w, T_i)$ refers to the frequency of w in T_i; $|T_i|$ is total number of keywords appeared in T_i; and $\alpha_i(\sum\alpha_i = 1)$ denotes the weight of T_i.

Given the confidence threshold η, the annotation keyword set of image I is:

$$Anno(I) = \{w|w \in WS_I \& Conf(w, I) \geq \eta\}. \qquad (21)$$

The training image set L_{train} is defined as those images whose semantic annotation keyword set are not empty, and the rest is the test image set, $L_{test} = L \setminus L_{train}$.

4.6 Annotation Algorithm

In this Subsection, we will present our CRF Model based Web image semantic annotation algorithm CRFM, which incorporate the keyword correlation, the textual and visual features of Web image in a unified graph model. The annotation algorithm is:

Algorithm 1 CRFM

```
Input: unlabeled image $I$, keywords correlation matrix $KCM$, keywords vocabulary $KV$, the number of annotation keywords $k$.
Output: Annotation keywords set $An$.
Initialize $An_0 = \emptyset$
for i = 1, 2, . . . , k do
    for each keyword $w \in KV$ do
        Construct the conditional random field for $w$, and estimate the parameters
        Calculate the probability $P(w|I)$ using Eqn.4
    end for
    Calculate $w^* = argmaxP(w|I)$
    Let $An_i = w^* \cup An_{i-1}$
end for
```

5. EXPERIMENTS

5.1 Experiment Setup

We downloaded the images and the accompanying Web pages by feeding the query keywords into Yahoo search engine, and parsed the html documents into DOM tree before extracting the embedded images and their corresponding associated texts to form the data set L. After parsing the pages and filtering the noisy images (such as the small logo images, the images with non-proper length/width ratio, etc.), we obtained the final set L with about 5,000 images.

We automatically selects a subset of about 1,000 images from L as training L_{train} using the proposed training set auto-generation method. The rest is used as test set L_{test}. We manually label the test images by 3 students (two of them are not familiar with this filed), and each image is labeled with 1-7 keywords. The vocabulary of manual annotations consists of about 137 keywords. Each image of L is segmented into 36 blobs based on fixed size grid, and 528 dimensional visual feature for each blob is extracted according to $MPEG7$ standard. Each image associates 5 types of associated texts: image file name, ALT texts (ALT tag), caption texts (Heading tag), surrounding texts and page title.

We further partitioned half of the training set as validation set to determine the model parameters, such as the smoothing parameter α and λ, and the weight shrinkage factor ρ. The corresponding values are set as 0.6, 0.6 and 0.8 respectively. In the auto-generation of training set, m is set to 4 (surrounding text is not included), $\alpha_i = 0.25(i = 1, \ldots, 4)$, and the confidence threshold η is set as 0.2. The *recall*, *precision* and F_1 measures are adopted to evaluate the performance in our experiments. That is, given a keyword w, let $|W_G|$ denote the number of human annotated images with label w in the test set, $|W_M|$ denote the number of images annotated with the same label by our algorithm. Then *recall*, *precision* and *F1* are respectively defined as: $Recall = \frac{|W_M \cap W_G|}{|W_G|}$, $Precision = \frac{|W_M \cap W_G|}{|W_M|}$, $F_1 = \frac{2(Precision \times Recall)}{Precision + Recall}$. The size of annotation is set to 5,

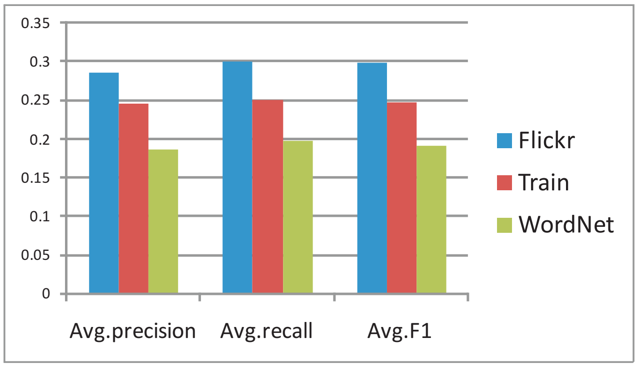


Figure 4: The effectiveness of different keyword correlation methods

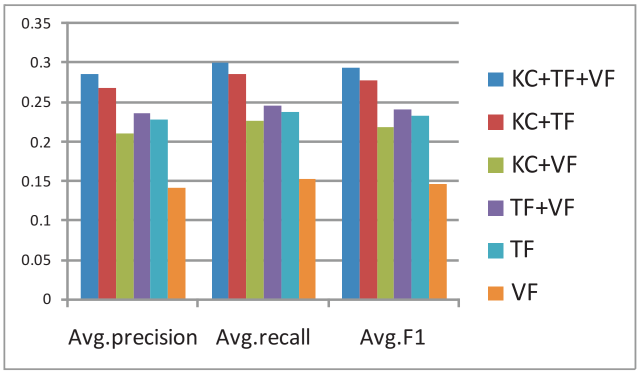


Figure 5: The effectiveness of different types of features

and the average *recall*, *precision* and *F1* over all keywords are calculated as evaluation of the overall performance.

5.2 Experimental Results

In our experiments, two baseline methods are used for comparison: (1) ModelAdp [28]: the baseline approach that do not use the keywords correlation and adaptively learn the textual model by Piecewise Penalty Weighted Regression model; and (2) ModelFMD: the baseline method that do not use the keywords correlation and learn a fixed textual model for estimating the semantic from the associated texts of Web image in the training stage.

5.2.1 *The Effectiveness of Keyword Correlation Matrix Derived From Flickr*

To test the effectiveness of the Keyword Correlation Matrix derived from Flickr (Flickr) proposed in this paper, we compare it with two methods: the training data based keywords correlation (Train) [30] and the WordNet-based keywords correlation (WordNet) [16]. We incorporate these three keyword correlations into our CRF based annotation framework, and compare their annotation performance. Figure 4 gives the comparison results.

It can be seen from Figure 4 that WordNet-based keywords correlation ("WordNet") performs the worst among the three methods, where it achieves the precision and recall of only 18.6% and 19.7% respectively. There are two main reasons for the poor performance of WordNet-based method. The first is that the similarities between annotations only depend on WordNet, which may not be proper for image annotation problem. There are 24 out of 137 words of the dataset that either do not exist in WordNet lexicon or have zero similarity with all other keywords. Moreover, the similarity defined using WordNet is sometimes not appropriate for the image annotation problem because it is defined using the wrong context. For example, "mountain" and "sky" usually appear in a scenery photo together, while "tree" and "flag" seldom simultaneously appear in an image. However the similarities in WordNet for the above two pairs of words are 0.1 and 0.1667 respectively, which is unreasonable.

Second, the training-based keyword semantic correction method ("Train") outperforms that using "WordNet" by a large margin. "Train" achieves the precision and recall of 24.6% and 25.1% respectively. However, the "Train" method has the limitation that its keyword similarity measurement depend only on the co-occurrence of keywords in the limited training data. This has resulted in poor generalization of the keyword correlations that it generates.

As expected, the Flickr-based keyword correlation ("Flickr") has the highest precision and recall of 28.6% and 30.1% respectively among the three methods. These is because it fully leverages the vast amount of manual image tagging resource, which ensures the effectiveness of keyword correlations generated for the image analysis problem.

5.2.2 *The Effectiveness of Different types of Features*

Our annotation approach mainly incorporates three type features: the Keywords Correlation (KC), the Textual Features (TF) and the Visual Features (VF). To test their contributions to the annotation performance, we compare the annotation performances of different combinations of features, Figure 5 presents the comparison results.

From Figure 5, we can draw the following conclusions: (a) "KC+TF+VF" has the highest precision and recall approaching 28.6% and 30.1% respectively. This shows that our annotation framework is able to integrate all types of features well for Web image annotation. (b) The "TF" and "KC+TF" outperform the "VF" and "KC+VF" respectively. This shows that the textual features are more effective than the visual features for image semantic annotation. (c) The "KC+TF+VF", "KC+TF" and "KC+VF" outperform the "TF+VF","TF" and "VF" respectively. This demonstrates that the keywords correlation can improve the performance of image semantic annotation.

5.2.3 *The Overall Performance of CRFM Algorithm*

Figure 6 compares the annotation performance of CRFM algorithm proposed in this paper with the baseline approaches. From the Figure, we can see that our proposed CRFM greatly outperforms both of the ModelAdp and ModelFMD. The performance is 29.3% for CRFM in terms of F_1 measure as compared to 23.6% and 21.6% for ModelAdp and ModelFMD respectively. This demonstrates that our annotation framework can incorporate the different types of information associated with Web image effectively, which results in great improvement of the annotation performance.

6. CONCLUSIONS

The growing Social Media resources, such as Flickr, Wiki etc. provide new opportunities for multimedia community

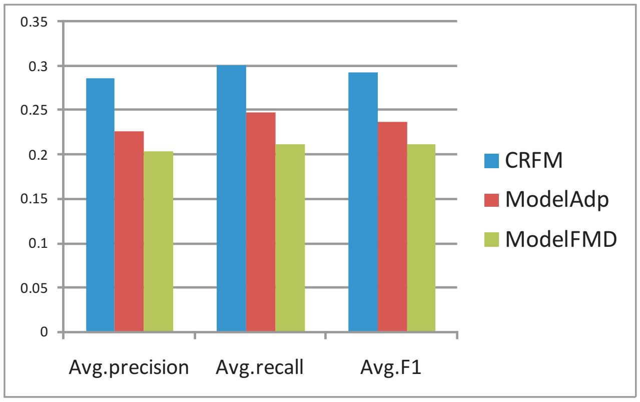


Figure 6: The overall performance of CRFM algorithm

to bridge the semantic gap. In this paper, we first explore the Flickr's Related Tags to derive a semantic correlation matrix. Then we demonstrate that the pure keyword correlation matrix derived from Flickr can be applied to improve the performance of Web image Annotation with our proposed CRF based annotation approach. The experimental results on the real Web image data set demonstrate the effectiveness of the proposed keyword correlation matrix and the Web image annotation approach. For future work, we plan to integrate the visual clues into the pure keyword space to further improve the effectiveness of the semantic correlation matrix.

Acknowledgments

This work was partially supported by the NSFC under Grant No.60403018, No.60773077 and No.90818023

7. REFERENCES

[1] http://www.flickr.net.

[2] E. Chang and et al. Cbsa: Content-based soft annotation for multimodal image retrieval using bayes point machines. *CirSysVideo*, 13(1):26–38, 2003.

[3] F. Christiane. Wordnet: An electronic lexical database. *MIT press*, 1998.

[4] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *ECCV*, pages 97–112, 2002.

[5] H. Feng, R. Shi, and T.-S. Chua. A bootstrapping framework for annotating and retrieving www images. *ACM Multimedia*, pages 960–967, 2004.

[6] S. Feng and R. Manmatha. A discrete direct retrieval model for image and video retrieval. *CIVR*, 2008.

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

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

[9] R. Jin, J. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. *ACM Multimedia*, 2004.

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

[11] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, 2001.

[12] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. *NIPS*, 2003.

[13] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP*, pages 1–19, 2006.

[14] X. Li, L. Chen, L. Zhang, F. Lin, and W. Ma. Image annotation by large-scale content-based image retrieval. *ACM Multimedia*, pages 607–610, 2006.

[15] J. Liu, M. Li, W. Ma, Q. Liu, and H. Lu. An adaptive graph model for automatic image annotation. *ACM Multimedia*, 2006.

[16] R. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. *AAAI*, 2004.

[17] B. Ricardo and R. Berthier. Modern information retrieval. *New York:ACM Press*, 1999.

[18] H. Sanderson and M. Dunlop. Image retrieval by hypertext links. *SIGIR*, pages 296–303, 1997.

[19] P. Schmitz. Inducing ontology from flickr tags. *WWW*, 2006.

[20] Y. Song, Z. Zhuang, H. Li, and Q. Zhao. Real-time automatic tag recommendation. *SIGIR*, pages 515–522, 2008.

[21] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan. Exploiting ontologies for automatic image annotation. *SIGIR*, pages 552–558, 2005.

[22] J. Tang, X.-S. Hua, G.-J. Qi, M. Wang, T. Mei, and X. Wu. Structure-sensitive manifold ranking for video concept detection. *ACM Multimedia*, 2007.

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

[24] V. Tseng, J. Su, B. Wang, and Y. Lin. Web image annotation by fusing visual features and textual information. *SAC*, pages 1056–1060, 2007.

[25] B. Wang, Z. Li, N. Yu, and M. Li. Image annotation in a progressive way. *ICME*, pages 1483–1490, 2007.

[26] P. Wang and C. Domeniconi. Building semantic kernels for text classification using wikipedia. *KDD*, 2008.

[27] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. *ACM MM*, 2008.

[28] H. Xu, X. Zhou, and L. Lin. Wisa: A novel web image semantic analysis system. *SIGIR*, 2008.

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

[30] 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.