Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Image interpretation: mining the visible and syntactic correlation of annotated words*

Ding-yin XIA†, Fei WU†‡, Wen-hao LIU, Han-wang ZHANG
(*School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)
†E-mail: {xiady, wufei}@cs.zju.edu.cn

Received Dec. 11, 2008; Revision accepted July 13, 2009; Crosschecked Oct. 18, 2009

Abstract: Automatic web image annotation is a practical and effective way for both web image retrieval and image understanding. However, current annotation techniques make no further investigation of the statement-level syntactic correlation among the annotated words, therefore making it very difficult to render natural language interpretation for images such as "pandas eat bamboo". In this paper, we propose an approach to interpret image semantics through mining the visible and textual information hidden in images. This approach mainly consists of two parts: first the annotated words of target images are ranked according to two factors, namely the visual correlation and the pairwise co-occurrence; then the statement-level syntactic correlation among annotated words is explored and natural language interpretation for the target image is obtained. Experiments conducted on real-world web images show the effectiveness of the proposed approach.

Key words: Web image annotation, Visibility, Pairwise co-occurrence, Natural language interpretation
doi:10.1631/jzus.A0820856 **Document code:** A **CLC number:** TP37; TP391

INTRODUCTION

With the development of the Internet, the number of digital images and videos has been growing rapidly. Annotation is an important way to succinctly describe the semantics of web images and it is very convenient for users to find similar images using annotated words. Many approaches have been proposed for automatic image annotation, such as machine translation (Duygulu *et al*., 2002), the statistical model (Li and Wang, 2006), the latent Dirichlet allocation model (Blei *et al*., 2003), maximum entropy (Jeon and Manmatha, 2004), the relevance model (Jeon *et al*., 2003), the coherent language model (Jin *et al*., 2004), inference networks (Metzler and Manmatha, 2004), and the tensor space model (Liu *et al*., 2008; Liu and Wu, 2009). All these methods however suffer from two main problems (Wang *et al*., 2008).

One is the well-known 'semantic gap' (Datta *et al*., 2008), since textual information around the images, such as surrounding text, unified resource locators (URLs) and title, is not fully utilized. Textual information around images has been used as approximate annotations by commercial image search engines, such as Google, Yahoo and Baidu. However, the correlations between images and words in the same web page are not fully discovered by these image search engines. Therefore, the annotations are very noisy, because many irrelative words are extracted. Furthermore, in most cases, words in the hosting web pages do not thoroughly describe the image within the same web page, so the annotations are incomplete and need to be extended. It is consequently necessary to re-rank all the imprecise annotations and remove the noisy ones by combining the visual content and semantic information.

Another important issue is the annotated words. Since 'a picture is worth a thousand keywords', we argue that the interpretation of images is very useful

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (Nos. 60533090 and 60603096) and the National High-Tech Research and Development Program (863) of China (No. 2006AA 010107)

for question answering (QA) services. Most current QA systems generally adopt one of the following three approaches (Pehcevski and Thom, 2007): natural language processing (NLP), information retrieval (IR), and template matching. However, these QA systems are based on text alone and can be difficult to use when users put forward questions about images with distinctive visual attributes.

This paper aims to interpret images from the accurately annotated keywords. The work similar to our idea is text-to-image synthesis (Zhu *et al*., 2007b) and photo-based question answering (Yeh *et al*., 2008). Text-to-image synthesis could generate pictures from general, unrestricted natural language text to convey the gist of the text. Photo-based QA supports direct use of photos in phrasing questions and finding answers.

In the well-known photo sharing website Flickr, there are more than 1.3×10^8 tags delivering almost 6×10^7 concepts (Wu *et al*., 2008) and each image is annotated by approximately 5~8 keywords. Although the linguistic correlation among annotated keywords could be used to interpret images, current annotation techniques make no further investigation of the statement-level (i.e., sentence level) syntactic correlation among the annotated words. Therefore, it is very difficult to render natural language interpretation for one image such as 'pandas eat bamboo' or 'a train gets across a bridge'.

In this paper, we propose an approach for automatic image interpretation with natural language. The idea in this approach mainly consists of two parts, namely first annotating the image and then generating its interpretation. Results of experiments conducted on real-world web images show the effectiveness of the proposed approach.

We summarize our main contributions as follows:

1. Given a web image and its surrounding text, we define the visibility of each word. Visibility of a certain word is the probability whether a word could be perceived visually. The textural and visible information of a word from the surrounding text are both considered for annotation.

2. After obtaining the initial annotation, we further mine the correlation of the annotation candidate keyword and the target image as well as the correlation of the pair of annotation candidate keywords to filter out the effect caused by noise.

3. The syntactic correlation between pairs of annotated words is analyzed to obtain the interpretation sentence of target images. We believe that the linguistic syntactic correlation could yield more interpretation for images than the few annotation words listed.

THE PROPOSED METHOD

Fig.1 outlines the overflow of our approach. For each potential annotated candidate, we obtain an image-word correlation matrix by the combination of tf*idf and visibility. Then we use latent textual and visual analysis to extend annotation keywords (Xia *et al*., 2008b). The final annotation words are ranked by visual correlation discovery and pairwise co-occurrence. At last, linguistic syntactic analysis is used to generate image interpretation.

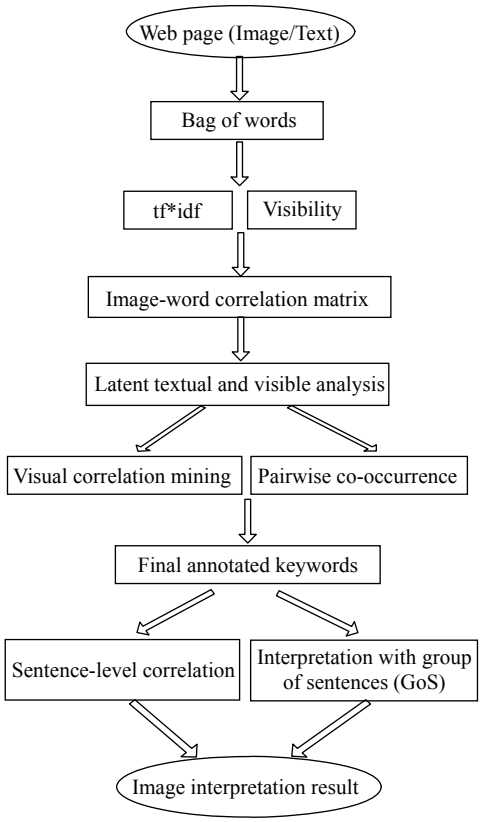

Fig.1 The flowchart of image interpretation

Image-word correlation matrix

Given a web image and its surrounding text, two factors always play an important role during annotation. One is textual information. After stop word removal and stemming are performed, the tf*idf value of each word is computed. The other is the visibility of words, defined as the extent to which a word can be perceived visually in this image. A word is visible when a good image can be drawn or found from it. On the contrary, a word is non-picturable when lacking an associated image that is clearly recognizable.

We use the WordNet (Miller, 1995) and Eq.(1) to evaluate the visibility of a word (or noun phrase) w_i. We manually select a set of landmark words which are top-k hot query words from Sogou (Liu *et al*., 2007) search engine. We try to cover the wide range of topics that are likely to arise in the subsequent image annotation. The visibility is set between 1 (visual) and 0 (not visual) according to Deschacht and Moens (2007), and the visibility of any word w_i outside of landmark words is given by

$$\mathrm{vis}(w_i)=\sum_j\left(\mathrm{vis}(w_j)\frac{\mathrm{sim}(w_i,w_j)}{\sum_j\mathrm{sim}(w_i,w_j)}\right), \quad (1)$$

where vis(w_i) is a value denoting the visibility of a word w_i, w_j is the landmark word whose visibility is set manually according to Deschacht and Moens (2007), and sim(w_i, w_j) is the similarity between word w_i and landmark word w_j, which is calculated by the least common subsumer (LCS) (Pedersen *et al*., 2004).

Current image search engines, such as Google, Baidu, and Yahoo, tend to utilize only textual information of web pages to index web images. Hence such information is very noisy (Fergus *et al*., 2005), and about 85% of returned images may be unrelated to the textual query when only tf-idf textual information is used. Therefore, the image-word correlation matrix is given by

$$\mathrm{cor}(w_i,I_j)=\mathrm{tf}_{i,j}\cdot\mathrm{idf}_i\cdot\mathrm{vis}(w_i), \quad (2)$$

where w_i is a word, I_j is an image, $\mathrm{tf}_{i,j}$ is the number of the words w_i associated with the image I_j, idf_i is a measure of the general importance of the word w_i, and vis(w_i) is the probability that the word w_i can be perceived in images, calculated as Eq.(1).

From Eq.(2), the word-image matrix consists of both textual and visible information, while the correlation cannot fully depict the hidden semantics of images in most cases. Hence, latent semantic analysis (LSA) (Deerwester *et al*., 1990; Cilibrasi and Vitanyi, 2006) is used to find the synonyms of annotation candidates. LSA can satisfy above requirements because of the way it adds more relevant words to the initial annotations and thus provides a better cover of annotations.

Annotation by visual correlation and pairwise co-occurrence

After the image-word correlation matrix is ready, we could obtain several candidate annotations. These candidate annotation terms are used as queries to find similar images, and data mining techniques can be adopted to de-noise and figure out salient terms or phrases from the search results to annotate the image. In order to obtain a final annotation, we discuss in this subsection how to estimate the visual correlation by pseudo-relevance-feedback (PRFB) (Yan *et al*., 2003) from the search result and the pairwise co-occurrence of annotated keywords. The algorithm is described in detail in Wu *et al*.(2009).

1. Visual correlation discovery by landmark affinity propagation clustering (Frey and Dueck 2007; Xia *et al*., 2008a)

Assuming w_i is a candidate annotation of target image I_j, the correlation between w_i and I_j is calculated by the similarity between I_j and the top k image cluster centers as follows:

$$\mathrm{rank}\,\boldsymbol{v}(w_i,I_j)=\frac{1}{k}\sum_{c_l\in C}\alpha_l\exp\left(-\frac{\mathrm{dis}(\boldsymbol{V}_j,\boldsymbol{V}_{c_l})}{\sigma}\right), \quad (3)$$

where w_i is a query word (also the candidate annotation keyword), and I_j is the target image; all of c_l are cluster centers in the set C. α_l is an adjustable parameter of each cluster center. dis($\boldsymbol{V}_j$, $\boldsymbol{V}_{c_l}$) is a certain distance metric between feature vectors $\boldsymbol{V}_j$ and $\boldsymbol{V}_{c_l}$, which is the Euclidean distance in our implementation.

2. Pairwise co-occurrence of annotated words

As stated in Wu *et al*.(2009), the textual ranking value for candidate annotation w_i is defined as the normalized summation of the textual similarities between w_i and the other words annotated in image I_j:

$$\text{rank}\,t(w_i, I_j) = \frac{1}{m}\sum_{w_j \in W_j} \beta_l \cdot \text{sim}\,t(w_i, w_j). \quad (4)$$

There are m words w_j in the annotation candidate set W_j for image I_j, β_l is a tuning parameter, and $\text{sim}\,t(x, y)$ is the pairwise co-occurrence of words x and y.

3. Combination of visual correlation and pairwise co-occurrence

After obtaining two different rank scores by Eqs.(3) and (4), we firstly normalize them into [0, 1] and then fuse them using a weighted linear combination scheme as follows:

$$\text{rank}\,(w_i, I_j) = \alpha \cdot \text{rank}\,v(w_i, I_j) + (1-\alpha) \cdot \text{rank}\,t(w_i, I_j), \quad (5)$$

where the weight α is between 0 and 1. Better performance is achieved when α is less than 0.5. This is because text features in web-based search engines are generally more effective than image features in our experiments.

IMAGE INTERPRETATION

After the annotated words are obtained, we intend to explore the sentence-level correlation of annotated words and to interpret images. Usually users want to know about what the image describes rather than independent annotated keywords alone. Therefore, interpreting each image with natural language is essential for image understanding and retrieval. Here we discuss how to generate natural language interpretation for images by mining knowledge from the Web.

Image interpretation by sentence-level correlation

In this subsection, we will introduce a simple version of image interpretation algorithm.

For one arbitrary annotated set Λ with m keywords, $\Lambda=\{w_i|1\le i\le m\}$, we need to figure out the meaningful sentences from the Web to describe images.

Here we consider only those sentences that contain exactly two keywords. This assumption does not impair the subtlety of the algorithm since the approach could be extended to find sentences consisting of more than two keywords. More discussion will be given later.

First the pairs of keywords with low semantic correlations in Λ are ruled out using Eq.(6) and we obtain a new keyword set Ψ:

$$\Psi = \{w_i \mid \text{NGD}(w_i, w_j) < \delta,\ i \ne j\}, \quad (6)$$

where δ is a threshold used to control the number of filtered out keywords, and $\text{NGD}(w_i, w_j)$ is the normalized Google distance between keywords w_i and w_j, as defined in Rui *et al*.(2007).

Given a pair of keywords in Ψ, to extract better sentences for the interpretation of an image, we define the concurrence frequency of keywords in the statement level rather than article/web page level as

$$F(\Psi \mid s) = P(w_i, w_j \mid s), \quad (7)$$

where s is an arbitrary sentence containing both w_i and w_j. Naturally, we choose only such sentences s that have large F to interpret the target image according to Eq.(8):

$$S = \{s_i \mid F(\Psi \mid s_i) > p\}, \quad (8)$$

where p is a threshold.

After we obtain the interpretation sentences for the target image, normalized Google generality defined in Eq.(9) is used to rank the candidate sentences for image interpretation:

$$\text{NGG}(s_i) = f(s_i) / N, \quad (9)$$

where $f(s_i)$ is the number of pages returned at searching s_i and N is the number of all web pages.

Image interpretation with group of sentences

By now we have discussed the proposed approach to interpret an image with a single sentence, and our experiment has substantiated its effect. However, one may argue that with the number of the annotation keywords larger than 2, the existence of the single sentences containing all the keywords remains unjustified. Moreover, it may be reasonable to imply that with an increasing number of annotation keywords, the number of qualified single sentences decreases rapidly to 0.

We have considered such potential risks of the approach in the previous subsection. Our further experiment showed that when we try to interpret an image with three annotation keywords with the above proposed method, in many cases the single sentences obtained are, as one may expect, not about a common sense or knowledge, while we can still obtain many single sentences that contain all the three keywords. Instead, they are about some specific stories, which may be even semantically irrelevant to the target image.

To understand such phenomena, we could turn to the habit of human being in describing an image or a scenario. For an image with simple high-level semantics containing few objects, one can always give his/her description with a simple sentence. However, when faced with an image containing an enormous number of objects with complicated semantics, people usually intend to describe them with several logically consecutive sentences. This may serve as a reasonable explanation for the problem mentioned above.

Moreover, in situations where the image to be described involves more than one story, which is quite common, one would always as well as have to interpret the image with more than one group of sentences with diverse semantics. Dealing with such cases, one must take into account the semantic diversity of sentences that he/she employs for the image interpreting task, to fully explore the semantics of an image.

To address the problem concerning a large number of annotation keywords and the semantic diversity of the final interpretation, we propose another approach for image interpretation: group of sentences (GoS). This method contains three phases: query generation, candidate GoS mining, and ranking, which are described in the following.

1. Query generation

In this phase, all annotation keywords are analyzed to find the most representative and important combinations of the keywords as the query to the Internet, in order to retrieve relevant documents, on which the following candidate GoS mining algorithm is applied.

As discussed before, the image annotation phase results in a set of keywords, Λ. According to our assumption, some of the keywords can be ruled out by the semantic correlation analysis, leaving two keywords as the query. Here we eliminate such a restraint and calculate the pairwise correlation matrix $\boldsymbol{C}$ of annotation keywords as

$$C_{ij} = \mathrm{NGD}(w_i, w_j), \tag{10}$$

where w_i, $w_j \in \Lambda$. Then the mean distance of each keyword to all the other keywords is computed by

$$\mathrm{MD}(w_i) = \frac{1}{r}\sum_{j=1}^{r} C_{ij}, \tag{11}$$

where $r=|C_{ij}|$ and $i \neq j$. Thus, we can determine the importance of an annotation keyword by the mean value of its distances to the other keywords ('mean distance' for short), and we select the top k keywords with smallest mean distances as the query keywords and denote the set of these keywords as Z.

However, one should neither query the Web with an input query obtained by simply combining all the keywords in Z, nor query the Web with each of the keywords in Z, since the first approach may lead to sparse or even no results, while the second approach may completely lose the latent correlation among the keywords. To strike a balance, we suggest querying with each pair of keywords at a time, which makes the number of our queries be C_k^2 in sum; it means that we need to query the Web C_k^2 times.

We select the top-100 returned documents of the search engine for each query, and collect all the returned documents in a set, RDoc. Thus, a corpus is built up where all the ensuing phases will be performed.

2. Candidate GoS mining

With the set of retrieved documents RDoc, we can extract the GoS of our interest, potentially qualified for the image interpretation task. In this paper, the mining phase is accomplished by Algorithm 1. Here we denote the number of sentences in a given group of sentence GoS_i as $|\mathrm{GoS}_i|$.

Algorithm 1 Mining candidate GoS

Input: the retrieved documents that are relevant to the queries generated in query generation, RDoc.

Output: the set of all candidate groups of sentences, Δ.

```
1   ω→∅; Δ→∅;
2   for each τ∈RDoc
3     for each sentence π∈τ
4       if π contains w_i (w_i∈Z) and |ω|<ε
            /* a keyword exists in π */
5         add π to ω;
```

```
6                 if ω contains all w_i∈Z
7                     add ω to Δ;
8                     ω→∅;
9                 end if
10            end if
11            else if |ω|<ε−1
              /* no keyword exists */
12                add π to ω;
                  /* but |ω| does not exceed ε */
13                move on to the next π;
14                goto line 4;
15            end else
16            else    /* |ω| exceeds ε */
17                ω→∅;
18            end else
19        end for
20    end for
21    return Δ;
```

/* where ω is a temporary GoS string, τ is a document within RDoc, π is a sentence within τ, $\varnothing$ indicates an empty set, and ε denotes the upper bound of $|\mathrm{GoS}_i|$ (i=1, 2, 3, …), which is given by the user. */

In this algorithm, we analyze each document in RDoc as a string and attempt to find all of its substrings that contain every keyword in Z. Two levels of loops are included in the algorithm: the outer loop handles a document as a whole, while the inner loop processes sentences in the current document, determining whether each sentence is supposed to be added to some GoS and whether a GoS has been formed.

3. Ranking candidate GoS

Since we have obtained the set of candidate GoS (i.e., Δ), a ranking phase is needed to determine the GoS that can best interpret the target image.

As discussed before, a good interpretation of an image should be precise and compact and be embodied with a semantic diversity. In response to these requirements, we propose the method of absorbing random walks rewarding compactness (ARWRC).

Absorbing random walks (ARW) (Zhu *et al*., 2007a) is a modified random walks model, dedicated to improve the diversity of ranking results. The classic random walks method is shown in Eq.(12):

$$\boldsymbol{P}=\lambda\tilde{\boldsymbol{P}}+(1-\lambda)\mathbf{1}\boldsymbol{r}^{\mathrm{T}}, \tag{12}$$

where $\mathbf{1}$ is an all-1 vector, $\mathbf{1}\boldsymbol{r}^{\mathrm{T}}$ is the outer product, $\boldsymbol{P}$ is the transition matrix in the new stage, $\tilde{\boldsymbol{P}}$ is the transition matrix of the previous stage, and λ is a constant. In ARW, the transition matrix $\boldsymbol{P}$ is represented as

$$\boldsymbol{P}=\begin{pmatrix}\boldsymbol{I}_{\mathrm{d}} & \mathbf{0}\\ \boldsymbol{R} & \boldsymbol{Q}\end{pmatrix}, \tag{13}$$

where $\boldsymbol{I}_{\mathrm{d}}$ is the identity matrix, corresponding to ranked items, and submatrices $\boldsymbol{R}$ and $\boldsymbol{Q}$ relate to rows of unranked items. Once an item is ranked, the corresponding state is considered to be absorbed and will just stay in its own state without any move. It means that, whenever an item is ranked, we update the transition matrix $\boldsymbol{P}$ using Eq.(13).

In each stage of ARW, the item with the greatest expected number of visits, proved to be given by the fundamental matrix as is defined by Eq.(14) (Doyle and Snell, 1984), is ranked and absorbed.

$$\boldsymbol{N}=(\boldsymbol{I}_{\mathrm{d}}-\boldsymbol{Q})^{-1}, \tag{14}$$

where N_{ij} refers to the expected number of visits to state j when the random walks process started from state i. Thus, after rank$\boldsymbol{P}$ stages of absorption, all items of the GoS are ranked.

Now take a look at the compactness of the GoS in our ranking algorithm. In order to make the interpretation more compact and to rule out specific stories conceived as not suitable for our task, we need to reward the GoS with great compactness. To define the compactness, it is necessary to refer to the expectation of image interpretation once more. An interpretation can be considered good if and only if it is directly describing the content or the correlations among the objects in the target image or telling us a piece of general knowledge relevant to the content of the image. Both of these two situations require that the GoS used as an interpretation be concise and not introduce much information that is not directly connected with the target image.

Based on the above analysis, we define the compactness of GoS_i as

$$\mathrm{CPT}(\mathrm{GoS}_i)=\exp\left[\frac{\mathrm{csim}(\mathrm{GoS}_i)}{\mathrm{sratio}(\mathrm{GoS}_i)\times\mathrm{wratio}(\mathrm{GoS}_i)}\right], \tag{15}$$

where $\mathrm{csim}(\mathrm{GoS}_i)$ represents the cosine similarity between the tf*idf vector of GoS_i and that of the annotation keywords, and

$$\mathrm{sratio}(\mathrm{GoS}_i)=\left|\mathrm{GoS}_i\right|/\left|Z\right|, \tag{16}$$

$$\text{wratio}(\text{GoS}_i) = \text{out}(\text{GoS}_i)/\text{wordcount}(\text{GoS}_i), \quad (17)$$

where out(GoS_i) denotes the number of words appearing in GoS_i but not in Λ, and wordcount(GoS_i) represents the number of all words appearing in GoS_i.

We apply the compactness of GoS to the transition matrix $\boldsymbol{P}$, rewarding the GoS with high compactness. Thus, a new transition matrix $\boldsymbol{W}$ is obtained:

$$W_{ij} = P_{ij} \times \text{CPT}(\text{GoS}_i). \quad (18)$$

Normalizing $\boldsymbol{W}$ by columns we obtain the transition matrix of ARWRC.

EXPERIMENTS

Datasets

The test images were crawled from the Internet. We selected 200 hot keywords from the search log, and submitted them to Google and Yahoo image search engines. For each text query, about 100 top-ranked images and their hosting web pages were crawled to the local machine. We have developed a hypertext markup language (HTML) parser system to extract the textual information (including surrounding text, title, URLs, alt attribute, anchor text and filename) of every image. Different parts of the textual content were given different weights. Furthermore, the 200 text queries were selected as landmark words and the visibility of them is defined as follows:

$$\text{vis}(s_j) = \frac{G_{\text{i}}(x)/G_{\text{t}}(x)}{G_{\text{I}}/G_{\text{T}}}, \quad (19)$$

where x is one query (annotated candidate) word, $G_{\text{i}}(x)$ is the count of retrieved pages when word x is the query for the Google image search engine, and $G_{\text{t}}(x)$ is the count of retrieved pages when word x is the query for the Google text search engine. G_{I} and G_{T} are the total numbers of images and web pages, respectively, of the Google search engine. Then the visibility of other words can be calculated using Eq.(1).

For each image, features with 423 dimensions were extracted, including 256 dimensions of color histogram, 6 dimensions of color moments, 128 dimensions of color coherence, 15 dimensions of MSRSAR texture, 10 dimensions of Tamura coarseness texture, and 8 dimensions of Tamura directionality texture. The selection of image features is not the focus of this paper; therefore, all other global or local features and the corresponding distance measures could be applied in our framework. In the interpretation phase, a corpus is needed and hereby Wikipedia was employed.

Evaluation metric

Since the acquisition of the ground truth is very expensive, the performance was evaluated from the view of image retrieval. In the experimental evaluation, 200 query words were selected and the relevance of resulting images was manually labeled. Then we reported the average precision and recall over the 200 words to evaluate the performance. Similar to text information retrieval, the number of correctly annotated images is defined as Num_{c}, the number of all retrieved images as Num_{r}, and the number of all relevant images in datasets as Num_{a}. Thus, the precision, recall and F-measure are computed as follows:

$$\text{precision}(w_i) = \text{Num}_{\text{c}}/\text{Num}_{\text{r}}, \quad (20)$$

$$\text{recall}(w_i) = \text{Num}_{\text{c}}/\text{Num}_{\text{a}}, \quad (21)$$

$$F(w_i) = \frac{2 \times \text{precision}(w_i) \times \text{recall}(w_i)}{\text{precision}(w_i) + \text{recall}(w_i)}. \quad (22)$$

Experimental results

In Section 2, we introduce the visibility to generate the image-word correlation matrix. For each web image, top k annotated words were selected by using the traditional tf*idf model and the visibility model in Eq.(2), respectively. Fig.2 shows the average F-measure value on different sizes of datasets in these two models. We can see a significant

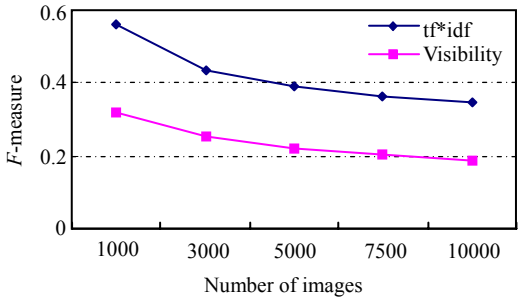

Fig.2 Average *F*-measure value comparisons on different sizes of datasets when combining textual and visible information taken during annotation

improvement over using tf-idf alone, even on the large scale datasets; hence, the combination of visible and textual information for web page annotation is more advantageous than textual information only.

Compared with the rank method using only visual information (visual) and the rank method using only textual information (textual), our automatic web image annotation approach achieves better performance, especially in retrieving more images. The main reason is that our approach considers both the visual correlation of images and the pairwise co-occurrence of words.

Table 1 shows parts of the annotation results using our approach. We queried Wikipedia for the top-100 relevant documents with the final annotation keywords.

Tables 2~4 display the interpretation results of the three images in Table 1 respectively. In each table, the interpretation results using single sentences with 2 and 3 annotation keywords, and the interpretation results using the GoS method with 3 annotation keywords are given.

Table 1 Parts of the annotation results

Image	Image No.	Annotations			
		After obtaining the image-word correlation matrix	After latent visual and semantic analysis	Final annotations (2 keywords)	Final annotations (3 keywords)
	1	panda, China	panda, China, tree, bamboo	panda, bamboo	panda, bamboo, China
	2	Jordan, NBA	Jordan, NBA, Bulls, basketball	Jordan, basket-ball	Jordan, basket-ball, NBA
	3	Clinton, President	Clinton, Democratic, President, Hillary	Clinton, Presi-dent	Clinton, Presi-dent, Democratic

Table 2 Part of interpretation results for Image 1 (panda) in Table 1

Keywords	Interpretation
panda, bamboo*	1. The Red Panda eats mostly bamboo. 2. Hulitherium may have fed on bamboo, like a panda. 3. Soft bamboo shoots, stems, and leaves are the major food source of the Giant Panda of China.
panda, bamboo, China*	1. Soft bamboo shoots, stems, and leaves are the major food source of the Giant Panda of China. 2. For example, the zoo raises 40 varieties of bamboo for the pandas on long-term loan from China, and it maintains 18 varieties of eucalyptus trees to feed its koalas. 3. Asian Black Bears share Giant Panda habitat in China's Wolong Reserve, where they feed occasionally, among other things, on bamboo, which is their more specialized relatives' favorite food.
panda, bamboo, China**	1. *Fargesia rufa* is a woody bamboo native to western China. It is known in Chinese as *qingchuan jianzhu*, meaning "*Qingchuan Fargesia*", *Qingchuan* being a county within the prefecture-level city of Guangyuan in the north of Sichuan. It is found at high elevations in the north of this province as well is in the south of Gansu. The plant is a significant source of food for the giant panda. 2. Though belonging to the order Carnivora, the Giant Panda has a diet which is 99% bamboo. The Giant Panda may eat other foods such as honey, eggs, fish, yams, shrub leaves, oranges, and bananas when available. The Giant Panda lives in a few mountain ranges in central China, in Sichuan, Shaanxi, and Gansu provinces. 3. Most famous of these is the giant panda, which survives in pockets of high-altitude bamboo forest across the southwest. Dhole in China is one of the least know species, it population in China are critically endangered.

* Single-sentence interpretation; ** GoS interpretation

Table 3 Part of interpretation results for Image 2 (Jordan) in Table 1

Keywords	Interpretation
Jordan, basketball*	1. Michael Jordan is an American basketball player. 2. For the basketball player see Jeffrey Jordan. 3. For the former professional basketball player, see Michael Jordan.
Jordan, basketball, NBA*	1. Jordan was drafted by the Seattle Supersonics in the 1993 NBA Draft after a storied basketball career at Kansas. 2. He is best known for his second book, *When Nothing Else Matters*, which chronicles basketball superstar Michael Jordan's last comeback to the NBA. 3. The book chronicles basketball superstar Michael Jordan's last comeback to the NBA playing for the Washington Wizards.
Jordan, basketball, NBA**	1. Michael Jordan was voted the Most Valuable Player of the series (he also had won the award the last five times the Bulls won the Finals: 1991, 1992, 1993, 1996, and 1997). This would be his sixth NBA championship and sixth Finals MVP award in six full basketball seasons, an unprecedented feat. 2. However in September 2001, Michael Jordan came out of retirement at age 38 to play basketball for the Washington Wizards. Jordan stated that he was returning "for the love of the game." Because of NBA rules, he had to divest himself of any ownership of the team. 3. He played high school basketball for Loyola Academy in Wilmette, Illinois. Jordan is the son of retired world champion NBA MVP Michael Jordan who played for the Chicago Bulls and Washington Wizards.

* Single-sentence interpretation; ** GoS interpretation

Table 4 Part of interpretation results for Image 3 (Clinton) in Table 1

Keywords	Interpretation
Clinton, president*	1. President Bill Clinton. 2. President Bill Clinton from 1993 to 1997. 3. Clinton won the 1992 presidential election.
Clinton, president, Democratic*	1. Presidential Election, Democratic President Bill Clinton received 72% of the Hispanic vote. 2. In 1996, Clinton became the first Democratic president to be reelected since Franklin D. 3. Future President Bill Clinton was the Democratic nominee for a seat in Arkansas, but lost.
Clinton, president, Democratic**	1. The caucus also re-nominated Clinton for a second term as Vice President. 2. His father, from the Democratic Republic of the Congo, named his son in honor of US president Bill Clinton years before the family was relocated by refugee agencies to Clarkston, Georgia. 3. The last time Louisiana went Democratic was for President Bill Clinton and Vice President Al Gore in 1996.

* Single-sentence interpretation; ** GoS interpretation

As the results showed, the sentence-level image interpretation method and the GoS method can be mutually complementary in a practical image interpretation task, where for annotations with few keywords, the former is qualified to render both precise and accurate interpretation, while the latter works better with annotations concerning a large number of keywords.

CONCLUSION

In this paper, we propose an approach for image interpretation by natural language via discovering both the visual correlation and pairwise co-occurrence. Extensive experiments on real-world web images verified the effectiveness and efficiency of the proposed framework.

We emphasize again that the interpretation of images with natural language could be used in many fields such as question answering, knowledge discovery and help of annotation since the linguistic statement-level correlation could be explored from structured or semi-structured corpora. In the future we would like to find more convinced linguistic models to study the hidden textual correlation for image interpretation.

References

Blei, D., Ng, A., Jordan, M., 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**(4-5):993-1022. [doi:10.1162/jmlr.2003.3.4-5.993]

Cilibrasi, R., Vitanyi, P., 2006. Automatic Extraction of Meaning from the Web. Proc. IEEE Int. Symp. on Information Theory, p.2309-2313.

Datta, R., Joshi, D., Li, J., Wang, J.Z., 2008. Image retrieval: ideas, influences, and trends of the new age. *ACM Comput. Surv.*, **40**(2): Article 5, p.1-60. [doi:10.1145/1348246.1348248]

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R., 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, **41**(6):391-407. [doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9]

Deschacht, K., Moens, M., 2007. Text Analysis for Automatic Image Annotation. 45th Annual Meeting Association for Computational Linguistics, p.1000-1007.

Doyle, P.G., Snell, J.L., 1984. Random Walks and Electric Networks. No. 22. Mathematical Association of America, Washington, D.C., USA.

Duygulu, P., Barnard, K., de Fretias, N., Forsyth, D., 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. Proc. European Conf. on Computer Vision, p.97-112. [doi:10.1007/3-540-47979-1_7]

Fergus, R., Li, F., Perona, P., Zisserman, A., 2005. Learning Object Categories from Google's Image Search. Tenth IEEE Int. Conf. on Computer Vision, p.1816-1823. [doi:10.1109/ICCV.2005.142]

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. *Science*, **315**(5814):972-976. [doi:10.1126/science.1136800]

Jeon, J., Manmatha, R., 2004. Using Maximum Entropy for Automatic Image Annotation. Proc. Int. Conf. on Image and Video Retrieval, p.24-32.

Jeon, J., Lavrenko, V., Manmatha, R., 2003. Automatic Image Annotation and Retrieval Using Cross-media Relevance Models. Proc. ACM SIGIR Conf., p.119-126. [doi:10.1145/860435.860459]

Jin, R., Chai, J.Y., Si, L., 2004. Effective Automatic Image Annotation via a Coherent Language Model and Active Learning. Proc. ACM Multimedia, p.892-899. [doi:10.1145/1027527.1027732]

Li, J., Wang, J.Z., 2006. Real-time Computerized Annotation of Pictures. Proc. ACM Multimedia, p.911-920. [doi:10.1145/1180639.1180841]

Liu, Y., Wu, F., 2009. Multi-modality video shot clustering with tensor representation. *Multim. Tools Appl.*, **41**(1):93-109. [doi:10.1007/s11042-008-0220-5]

Liu, Y., Fu, Y., Zhang, M., Ma, S., Ru, L., 2007. Automatic Search Engine Performance Evaluation with Click-through Data Analysis. Proc. 16th Int. Conf. on World Wide Web Conf., p.1133-1134. [doi:10.1145/1242572.1242731]

Liu, Y., Wu, F., Zhuang, Y., Xiao, J., 2008. Active Post-refined Multi-modality Video Semantic Concept Detection with Tensor Representation. Proc. ACM Multimedia, p.91-100. [doi:10.1145/1459359.1459372]

Metzler, D., Manmatha, R., 2004. An Inference Network Approach to Image Retrieval. Proc Int. Conf. on Image and Video Retrieval, p.42-50.

Miller, G.A., 1995. WordNet: a lexical database for English. *Commun. ACM*, **38**(11):39-41. [doi:10.1145/219717.219748]

Pedersen, T., Patwardhan, S., Michelizzi, J., 2004. WordNet-Similarity: Measuring the Relatedness of Concepts. Proc. 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics, p.38-41.

Pehcevski, J., Thom, J.A., 2007. Evaluating Focused Retrieval Tasks. SIGIR Workshop on Focused Retrieval, p.33-40.

Rui, X., Yu, N., Wang, T., Li, M., 2007. A Search-based Web Image Annotation Method. IEEE Int. Conf. on Multimedia and Expo, p.655-658. [doi:10.1109/ICME.2007.4284735]

Wang, J.Z., Geman, D., Luo, J., Gray, R.M., 2008. Real-world image annotation and retrieval: an introduction to the special section. *IEEE Trans. Pattern Anal. Mach. Intell.*, **30**(11):1873-1876. [doi:10.1109/TPAMI.2008.231]

Wu, F., Xia, D., Zhuang, Y., Zhang, H., Liu, W., 2009. Web Image Interpretation: Semi-supervised Mining Annotated Words. IEEE Int. Conf. on Multimedia and Expo, p.1512-1515. [doi:10.1109/ICME.2009.5202791]

Wu, L., Hua, X.S., Yu, N., Ma, W.Y., Li, S., 2008. Flickr Distance. Proc. ACM Multimedia, p.31-40. [doi:10.1145/1459359.1459364]

Xia, D., Wu, F., Zhang, X., Zhuang, Y., 2008a. Local and global approaches of affinity propagation clustering for large scale data. *J. Zhejiang Univ. Sci. A*, **9**(10):1373-1381. [doi:10.1631/jzus.A0720058]

Xia, D., Wu, F., Zhuang, Y., 2008b. Search-Based Automatic Web Image Annotation Using Latent Visual and Semantic Analysis. Pacific-Rim Conf. on Multimedia, p.842-845. [doi:10.1007/978-3-540-89796-5_95]

Yan, R., Hauptmann, A., Jin, R., 2003. Multimedia Search with Pseudo-relevance Feedback. Proc. Int. Conf. on Image and Video Retrieval, p.238-247.

Yeh, T., Lee, J.J., Darrell, T., 2008. Photo-based Question Answering. Proc. ACM Multimedia, p.389-398. [doi:10.1145/1459359.1459412]

Zhu, X., Goldberg, A.B., van Gael, J., Andrzejewski, D., 2007a. Improving Diversity in Ranking Using Absorbing Random Walks. Proc. 8th Annual Meeting of the North American Chapter of the Association for Computational Linguistics.

Zhu, X., Goldberg, A.B., Eldawy, M., Dyer, C.R., Strock, B., 2007b. A Text-to-picture Synthesis System for Augmenting Communication. Integrated Intelligence Track of the 22nd AAAI Conf. on Artificial Intelligence, p.1590-1595.