

High-level semantic image annotation based on hot Internet topics

Xiaoru Wang · Junping Du · Shuzhe Wu · Xu Li ·
Haiming Xin · Yu Zhang · Fu Li

Published online: 8 November 2013

© The Author(s) 2013. This article is published with open access at Springerlink.com

Abstract Images are complex multimedia data that contain rich semantic information. Currently, most of image annotation algorithms are only annotating the object semantics of images. There are still many challenges on high-level semantic image annotation. The major issues are the lack of effective modeling method for the high-level semantics of images and the lack of efficient dynamic update mechanism for the training set. To address these issues, we propose a high-level semantic annotation method based on hot Internet topics in this paper. There are two independent sub tasks in our method: dynamic update of the training set based on hot Internet topics and search-based image annotation. In the first sub task, we propose to model the abstract semantics of images based on three relationships: image-to-image similarity relationship, topic-to-topic co-occurrence relationship, and image-to-topic relevance relationship. Through the complex graph clustering, the hot Internet topics are extracted for images with consistent visual and semantic contents. Then the dynamic update mechanism will update the original training set with the new topics and images. It avoids the huge computing cost in traditional update methods and does not need to re-calculate the whole mapping relationship between the semantic concepts and visual features. In the second sub task, given a query image, it first searches for similar candidates in the annotated training set via visual features. Then the hypergraph modeling and spectral clustering are exploited to filter out the images with irrelevant semantics. The keywords will be extracted for annotation from the remaining images according to an annotation probability. Extensive experiments

X. Wang (✉) · J. Du · S. Wu · X. Li · H. Xin · Y. Zhang
Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing 100876, China
e-mail: wxr@bupt.edu.cn

F. Li
Department of Electrical and Computer Engineering, Portland State University,
Portland, OR, USA

have been conducted and the results demonstrate that our algorithm could achieve better annotation performance than the state-of-the-art algorithms. And the update mechanism could extend the training set efficiently so that the coverage of the semantics in the training set wouldn't be obsolete.

Keywords Abstract semantics · Complex graph · High-level semantics · Hot Internet topic · Hypergraph · Image annotation

1 Introduction

The number of images on the Internet has been growing explosively with the widespread use of smart phones, digital cameras, and other portable devices. With millions of photos uploaded every day, it has been a great challenge to effectively organize and retrieve the images on the Internet, which is an important research area with broad practical usages. Image annotation is a crucial technology that facilitates image retrieval by adding keywords to images. In general, new images are annotated by finding and building a relevance model that connects high-level semantics and low-level visual features of images. And the high-level semantics will become the keywords for an image. Thus, image annotation technology could transform the image retrieval problem into a more mature text retrieval problem. Users can use keywords to retrieve the images they require. If users provide a query image, the semantics of the query image can also be constructed by searching and processing the annotated images that are similar to the query image.

“A picture is worth a thousand words.” Images are a type of complex multimedia data that contain an abundance of semantic information. Eakins [8] proposed that the semantics of images have three levels. The first level is the low-level semantics, which are the low-level visual features extracted from images such as color and texture. The second level is the object semantics, which are the objects recognized in the image such as tiger, apple, etc. In general, the object categories are inferred from the extracted visual features. The third level is the abstract semantics, which is a higher level inference of the object semantics. Example words and phrases include those expressing concepts (e.g. geek, CEO, art), properties (e.g. nutrient, tasty), behaviors (e.g. resign, lose weight), etc. Currently, the majority of the annotation research focuses on the object semantics based on the visual content of images. There are still many challenges to address with regard to abstract semantics annotation. In this paper, we consider two critical problems with abstract semantics annotation.

The first problem is the lack of effective modeling methods for abstract semantics, which has given rise to a surge of interest in the extraction of precise keywords from the text or user comments associated with an image. Several algorithms have been proposed for automatically assigning keywords to images or image regions. They first select salient terms or more visible words from the associated text and calculate the frequency or co-occurrences of these words [6, 35]. Then an image is annotated using the words with the maximum annotation probability. In most cases, the annotation words selected using such a method represent the objects in an image, i.e., the object semantics it contains. It is still difficult to represent the abstract semantics. Thus when users submit query keywords for images, the effective words to use will be

constrained by the object semantics of the image. We believe that a more natural and personalized way of retrieving images is to submit query keywords expressing high-level semantics, e.g., a hot Internet topic. For example, there was a hot topic on the Internet about a food safety accident of red yolk stained with Sudan Red. If “Sudan Red” or “stain” are included as query keywords, users may want to retrieve images about the impact of Sudan on food. Only when abstract semantics are expressed in annotations can the related images be retrieved as desired. So, it is critical to find a way of modeling and extracting abstract semantics of images.

The second problem with existing annotation approaches is the lack of dynamic updating mechanism for the training set. Current approaches adopt machine learning techniques or other relevance modeling methods to learn from a static annotated training set and identify the keywords for new images based on the learned model [14]. This process can be viewed as mapping low-level features of images to high-level semantic concepts. Therefore the annotation results are restricted by the visual features in the training set and the semantics covered by the training set annotations. Without an effective updating mechanism, the training set only provides a fixed vocabulary for annotation and cannot grow to cover newly formed semantics (e.g. new events, temporal hot topics, etc.). Furthermore, regular update of the training set requires re-computing the mapping relationship between annotations and visual features, which can be extremely time-consuming and computationally intensive for large-scale training set.

To address the issues above, we propose a novel high-level semantic annotation method for images based on hot Internet topics. There are three main contributions in this paper. First, we propose to model high-level semantics based on hot Internet topics. We use Latent Dirichlet Allocation (LDA) to analyze the texts on the related web pages to build the topics. After that, three kinds of the relevance relationships are constructed: the topic-to-topic co-occurrence relationship, the topic-to-image relevance relationship and the image-to-image visual similarity relationship. Through the modeling and clustering of complex graph, the hot topics are formed by clustering the LDA topics of similar images, and the related images are annotated with top words in the corresponding hot topics. Second, we propose a dynamic update mechanism for the training set based on hot Internet topics. Through the discovery and tracking of hot Internet topics, representative keywords are selected to annotate related images in the training set. Note that the words for annotation are based on texts newly searched on the Internet so that the vocabulary for annotation can be constantly extended. In addition, our update mechanism does not require the re-computing of the whole mapping relationship between the annotations and low-level visual features of all images. We only update the annotations of the related images in the original training set. This will greatly reduce the update cost and make the extension of the training set much easier. So, the semantic coverage will be increased gradually and the dynamic update of the training set could be achieved, which are very critical for a good image retrieval system. Third, we propose a new search-based image annotation mechanism, which uses the hypergraph modeling and spectral clustering to filter out the semantically irrelevant images. Given a query image, we search for the candidates from the training set according to the visual similarity. A hypergraph is constructed for the candidates and several clusters could be formed via clustering process. Small clusters are considered to be outliers and are

discarded. The selected cluster will have both similar visual features and consistent semantics with the query image. So, the annotation results from this cluster will deliver better performance.

The paper is organized as follows. Section 2 discusses previous work on image annotation, which provides readers with a general idea of related work. Section 3 explains the basic design concept. Sections 4 and 5 present the approach in detail. Section 6 contains our evaluations and Section 7 presents our conclusions.

2 Related work

Research on image annotation technology has been conducted for many years. In general, researchers have proposed knowledge modeling methods for automatic annotation, such as classification-based methods [2, 4, 15, 23], graphical model-based methods [36, 37], cross-media modeling methods [9, 11, 17], and translation model-based [5, 7, 12] methods. The key attributes of these methods are using different machine learning algorithms on a training set and the construction of mapping relationships between the semantic concepts and low-level features. These mapping relationships are used to annotate new images. However, the size of the training sets used by these methods is limited, so their quality and performance will degrade when handling large-scale Internet data. This is because the coverage of the semantics in the training set is also limited and it cannot be updated in real-time. In most cases, the annotation results are simply the object semantics of the images.

As the wide use of the social network, more and more users could share their images with others on the Internet. They could also add comments to some images to express feelings, which is called “social tagging”. For example, these websites include Flickr, Photosig, Delicious, LabelMe, Peekaboom, etc. Social tagging suffers from two drawbacks, though it is easy to perform. First, the tags provided by Flickr users actually contain many noise [13]. Second, there is ambiguity in the user tagging. Users often use some common tags for different objects. So these common tags are semantically ambiguous. This is the reason why many similar images can't be retrieved based on keyword mechanism. Wu et al. [34] proposed a tag recommendation framework taking advantage of correlations between tags and visual content. Weak rankers are learned with this multi-modality model and combined using rank-boost algorithm. However, these methods usually need to build a static vocabulary according to the training set and calculate the mapping relationship among tags in the vocabulary. Once a new word or a new annotated image is added, the whole mapping relationship in the vocabulary should be re-calculated. This is a huge computing cost for a large-scale image database.

Recently, the research focus of image annotation has changed to large-scale Internet image annotation [28, 30]. Given a query image for annotation, its semantic contents can be extracted from similar images that have been annotated. Thus, if similar candidates can be retrieved from Internet, an annotation of the query image can be obtained [16, 24, 29, 31, 32]. In general, these methods require a query word when searching for similar images and they submit the query word to a text-based search engine. The key issue that affects these methods is the algorithm used to identify accurate query words. Zhang et al. [38] proposed an image annotation method that could acquire the initial query words automatically. The key concept in

this method was using CBIR technique to find similar candidates and their annotations in a database. Next, the initial query words were derived and used to search for similar images from the Internet after filtering out noisy words and performing a sort process. However, these methods just used a search based algorithm to retrieve similar images from the Internet instead of the training dataset, and then derived annotations from these retrieved images. They had no capability to annotate images based on hot Internet topics.

The data found on the Internet is a type of cross-media data, so the images and their associated text may have some forms of semantic relationships. This relationship may provide an effective method for image annotation. Zhu et al. [39] selected salient terms from the associated text for image annotation. Wu et al. [33] proposed more accurate annotations by calculating the contribution of each word to its visibility model. Monay and Gatica-Perez [21] introduced a latent topic model, probabilistic Latent Semantic Analysis (pLSA), into an image annotation algorithm and trained two pLSA models based on the visual features and associated text, respectively. This method then merged the topics in both models by assigning weights according to the entropy of the visual word distribution. These methods combined the relationships of different features and improved the annotation accuracy. However, these modeling approaches mainly considered the semantic relationships between images and salient words, and the relationships between salient words. In general, these salient words represent the objects in an image. Thus, assigning the salient words as annotations still fails to describe the high-level abstract semantics.

3 Basic design idea

In this paper, a novel high-level semantic image annotation algorithm based on hot Internet topics is proposed. As shown in Fig. 1, the algorithm consists of two sub tasks: search-based image annotation and dynamic update of the training set based on hot Internet topics. The former aims at annotating a given image based on the annotated set, while the latter prevents annotations in the set to be obsolete and keeps the set in synchronization with the latest hot topics on the Internet.

3.1 Dynamic update of the training set based on hot Internet topics

The volume of data on the Internet is huge and it grows continuously. To discover the hot topics in such a large scale data set, we follow the 4 steps below.

1. **Corpus collection:** For a specific accident or event, the title of it is used to search for related web pages using TBIR technique. Besides, web pages containing both images and texts are regularly collected to construct the corpus for hot topic discovery.
2. **Hot topic discovery:** To discover hot topics in the collected corpus, we first use LDA modeling [1] to learn the topic distributions in the text corpus. Then three kinds of relationships, including image-to-image, topic-to-topic and image-to-topic relevance, are constructed. We build a complex graph based on these relationships and perform a clustering operation on the constructed graph [18]. The complex graph takes images and LDA topics as vertices and is partitioned according to three sets of relationships. The mapping between the image clusters

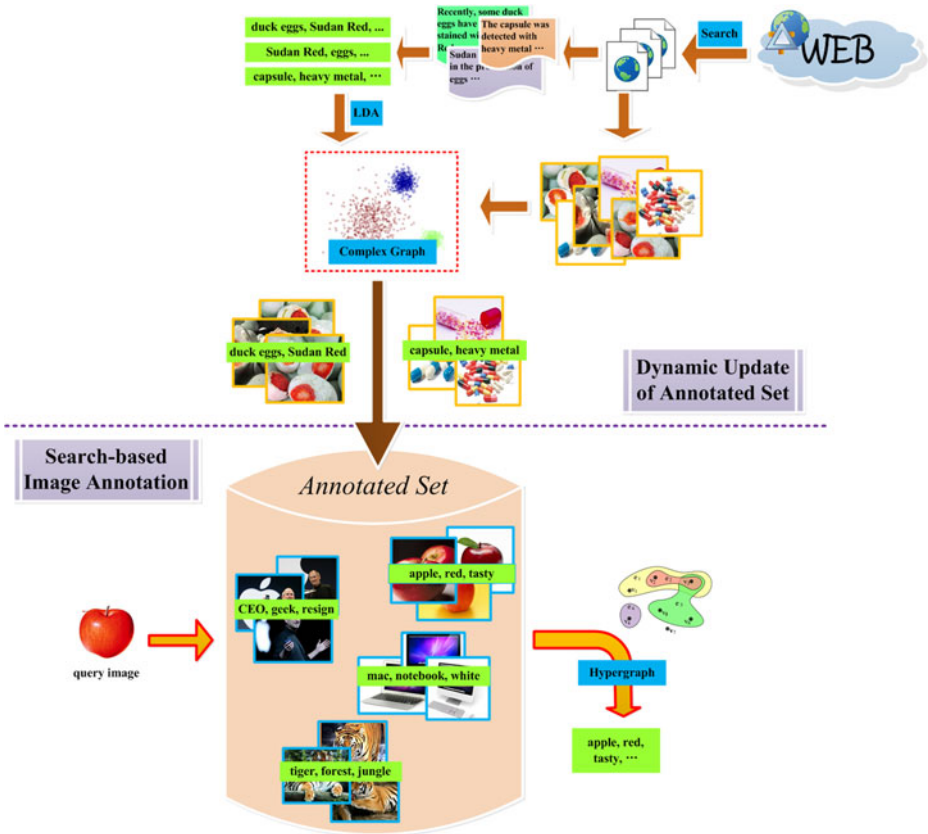


Fig. 1 Image annotation based on the two sub tasks

and the topic clusters is built via the complex graph clustering and the images in one cluster will have similar visual features and consistent semantics. In this way, sub hot topics could be formed even when the initial title is polysemous or ambiguous.

3. **Keyword selection:** With the hot topics discovered and the corresponding words representing them, χ^2 is used to select for each hot topic the keywords so that they can be understood by users. And the corresponding images are annotated using the selected keywords.
4. **Database update:** After the images are annotated, they are added to the training set, the semantics coverage of which is thus expanded and in synchronization with the latest hot topics on the Internet.

The core step is the second one. In this step, complex graph clustering is performed based on three sets of relationships, and hot topics are formed with establishment of the mapping relationship between obtained image and topic clusters. Note that an image is represented as a vector of extracted visual features, while a topic is represented as a distribution over text words. They belong to essentially different feature spaces and it is difficult to build a relevance relationship between them.

Existing algorithms generally assume that an image and its associated text are semantically relevant. Such relevance can bridge the different feature spaces and connect an image with a topic. However, because the text associated with a given image are extracted from the corresponding web page, which usually contains multiple images, some content of the text is actually not relevant to the given image at all. It is incorrect if the topics extracted from the irrelevant content are connected with the given image.

In addition, considering the polysemy of words [25], the visual content of the images found by the same query keyword can be quite different (e.g. The word “apple” indicates both the fruit apple and the Apple Inc.). Consequently, the images from one query keyword may be clustered into several sub categories.

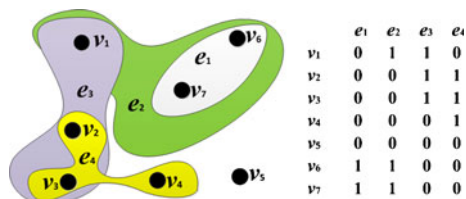
To overcome such problems and establish appropriate relevance between the image and topic clusters, the relevance between an image and a topic, the visual similarity between images, and the co-occurrences of topics are all taken into consideration, resulting in three sets of relationships. Complex graph provides a natural way of modeling both the images and the texts simultaneously to exploit the relationships because it allows multi-type vertices to be connected together. By clustering on the complex graph, images in the same cluster will not only be visually similar but also share the same hot topic.

3.2 Search-based image annotation

In this sub task, we will follow the 3 steps below to construct the annotations for a query image by leveraging the annotated training set.

1. The annotated set is searched for the images visually similar to the given one. Considering the existence of semantic gap problem, some of the found images will be semantically irrelevant. Therefore it is inappropriate to select keywords among their annotations directly.
2. To filter out the semantically irrelevant images, a hypergraph is constructed with the found images as vertices and their annotations as hyperedges. That is, if two images are annotated with the same word, then the corresponding two vertices are connected by a hyperedge. An example of hypergraph is shown in Fig. 2. With the hypergraph, spectral clustering is performed to partition it into multiple clusters, and the small clusters will be regarded as semantically irrelevant and discarded.
3. The cluster of images that are most relevant to the query image is identified among the remaining clusters according to visual similarity. Then for each annotation in the cluster, its relevance to the query image is computed. And the most relevant annotations will be used to annotate the query image.

Fig. 2 Hypergraph and matrix **H**



Note that hypergraph provides a natural and concise way of modeling the relationships between images based on their annotations by allowing multiple vertices to be connected by a single hyperedge. It can perfectly express that one annotation can be assigned to multiple images. And moreover, one vertex is also allowed to be included in different hyperedges, which describes that one image can have multiple annotations. Annotations are used to construct hyperedges because they represent the semantics contained in the images.

By partitioning the constructed hypergraph, images that are semantically irrelevant can be identified by considering the fact that their semantics will be quite different from that of the query image and vary among themselves. Therefore those images will only form small clusters, whereas others will form larger ones.

4 Dynamic update of the training set based on hot Internet topics

As Section 3 described, the dynamic update of the annotated training set consists of four steps. The core algorithm of discovering hot topics will be discussed in detail in this section and the pseudocode is shown in Algorithm 1.

Algorithm 1 Discovery of hot topics

- 1: **Input:** Extended image set V_e , associated text set T_e
 - 2: **Output:** Annotations of hot topics A
 - 3:
 - 4: $R_v \leftarrow \text{ComputeImageRelevance}(V_e)$
 - 5: $R_t \leftarrow \text{ComputeTopicRelevance}(T_e)$
 - 6: $p(z|d) \leftarrow \text{GibbsSamplingLDA}(T_e)$
 - 7: $R_{vt} \leftarrow \text{ComputeImageTopicRelevance}(p(z|d), R_v)$
 - 8: $G \leftarrow \text{ConstructComplexGraph}(R_v, R_t, R_{vt})$
 - 9: $S \leftarrow \text{ComplexGraphClustering}(G)$
 - 10: $A \leftarrow \text{SelectMostRelevantWords}(S)$
-

4.1 Image representation and similarity measurement

Three types of features are extracted to represent images in our experiment, i.e. color histogram [26], wavelet texture [20] and SIFT [19], the dimensions of which are 64, 128 and 500 respectively. Note that the choice of features for image representation is not the focus of our paper and in effect many other visual features can also be used as replacement. Specifically, color histograms are computed in the LAB color space. The lightness and two color components are all uniformly quantized into 4 bins, and the χ^2 distance is computed to measure the similarity between two images in terms of their color histograms. The texture feature vector of an image is computed with both pyramid- and tree-structured wavelet transform by decomposing, at different levels, the sub-bands obtained through filtering. And it consists of the means and standard deviations of all the energy distributions [3]. As for SIFT, a visual vocabulary of size 500 is constructed through k-means clustering. Euclidean distances are used to define the visual similarity between two images.

4.2 Building three relevance relationships

Three kinds of relevance relationships are considered in the image and text modeling, i.e. the image-to-image similarity relationship, the topic-to-topic co-occurrence relationship, and the image-to-topic relevance relationship.

The topics are extracted from the associated text set T_e using Latent Dirichlet Allocation (LDA) [1]. LDA is a generative model for collections of discrete data such as text corpora, and it is endowed with three layers corresponding to documents, topics, and words respectively. A document is regarded as a mixture of an underlying set of topics. This provides a representation of documents as topic distributions, allowing them to be analyzed effectively in the latent topic space that is usually a much lower dimensional one. Gibbs sampling is adopted for parameter estimation during LDA modeling [10], after which each word is assigned a topic label and the topic-document distribution $p(\mathbf{z}|d_j)$ can be determined. Equation (1) shows how the topic-document distributions can be estimated.

$$p(z_k|d_j) = \frac{n_k^{(d_j)} + \alpha}{n_{\bullet}^{(d_j)} + T\alpha} \quad (1)$$

- α is the Dirichlet hyperparameter for topic-document distributions
- T is the number of topics
- $n_k^{(d_j)}$ is the frequency of topic k in document d_j
- $n_{\bullet}^{(d_j)}$ is the total frequency of all topics in document d_j

Image-to-image similarity relationship According to Section 4.1, a similarity matrix R_v can be constructed for the extended image set V_e , whose element $R_v(i, j)$ represent the similarity between image i and j in the set.

Topic-to-topic co-occurrence relationship The relevance relationship between topics can be constructed based on the LDA topic assignments of each word in all the documents. Let R_t be the topic co-occurrence matrix. Its element $R_t(i, j)$ is defined in (2).

$$R_t(i, j) = \frac{C(z_i \cap z_j)}{C(z_i \cap z_j) + C(\bar{z}_i \cap \bar{z}_j)} \cdot \frac{C(z_i \cap z_j)}{C(z_j)} + \frac{C(\bar{z}_i \cap \bar{z}_j)}{C(z_i \cap z_j) + C(\bar{z}_i \cap \bar{z}_j)} \cdot \frac{C(\bar{z}_i \cap \bar{z}_j)}{C(\bar{z}_j)} \quad (2)$$

$C(z_i \cap z_j)$ is the co-occurrence count of topics z_i and z_j , i.e. the number of times that both topics are assigned to some words in the same document, and $C(\bar{z}_i \cap \bar{z}_j)$ is the neither-existence count of topics z_i and z_j , i.e. the number of times that neither of the two topics are assigned to any words in the same document.

Image-to-topic relevance relationship The image-to-topic relevance relationship can be measured using the conditional probability of the topic given the image, i.e.

$p(z_j|I_i)$, which can be decomposed by considering all the images similar to the given one as in (3).

$$\begin{aligned}
 p(z_j|I_i) &= \sum_{I_{sim} \in V_s} p(z_j|I_{sim})p(I_{sim}|I_i) \\
 p(z_j|I_{sim}) &\propto p(z_j|d_{sim}), \text{ learned from LDA} \\
 p(I_{sim}|I_i) &\propto \text{similarity}(I_i, I_{sim})
 \end{aligned}
 \tag{3}$$

The decomposition leads a helpful intuition that if images similar to the given one are relevant to a topic, then the given one should also be relevant to it. To avoid iterative computation to reach the convergence with such a recursive definition, the corresponding text is used to stand for the image, and the topic-document distributions learned with LDA is employed to calculate the relevance values needed. The visual similarity can be viewed as a weight for combination, ensuring that more similar images make greater contributions.

4.3 Modeling the three relevance relationships using a complex graph

To represent and exploit the three relationships in an unified framework, a complex graph $G = \{V_1, V_2, E\}$ is constructed, which provides a natural way of modeling images and LDA topics simultaneously by allowing multi-type vertices to be connected [18].

The vertex set V_1 and V_2 correspond to LDA topics and images respectively, and the edge set E contains connections between vertices including those between homogeneous vertices and those between heterogeneous ones. The edge set can be written as $E = \left\{ \left\{ S \in R_+^{|V_1| \times |V_1|} \right\}, \left\{ A \in R_+^{|V_1| \times |V_2|} \right\} \right\}$, where S represents the weights of the homogeneous edges connecting vertices in V_1 (see Eq. (2)), and A represents the weights of the heterogeneous edges connecting vertices in V_1 with those in V_2 (see Eq. (3)). Based on the complex graph constructed, the images and LDA topics can be separately clustered while being constrained by each other. Mapping relationship between the obtained image and LDA topic clusters are established according to the three kinds of relevance relationships.

The complex graph is partitioned to optimize the objective function L defined in (4) [18].

$$\begin{aligned}
 &\arg \min_{\mathbf{C}^{(1)}, \mathbf{C}^{(2)}} L \\
 L &= \|S - \mathbf{C}^{(1)} \mathbf{D} (\mathbf{C}^{(1)})^T\|^2 + \|A - \mathbf{C}^{(1)} \mathbf{B} (\mathbf{C}^{(2)})^T\|^2 \\
 \text{s.t. } &\mathbf{C}^{(1)} \in \{0, 1\}^{|V_1| \times K_1}, \mathbf{C}^{(2)} \in \{0, 1\}^{|V_2| \times K_2}
 \end{aligned}
 \tag{4}$$

$\mathbf{C}^{(1)}$ denotes the cluster membership matrix for the vertices in V_1 , and $C_{ij}^{(1)}$ is the weight between vertex i and cluster j in V_1 . $\mathbf{C}^{(2)}$ denotes the cluster membership matrix for the vertices in V_2 , and $C_{ij}^{(2)}$ is the weight between vertex i and cluster j in V_2 . The inter-type pattern matrix \mathbf{B} denotes the link patterns between the vertices in V_1 and those in V_2 , and $B(i, j)$ is the link strength between cluster i in V_1 and cluster j in V_2 . The intra-type cluster pattern matrix \mathbf{D} denotes the link patterns in the same type of vertices, and $D(i, j)$ is the link strength between cluster i and cluster j in V_1 . In general, the matrices \mathbf{D} and \mathbf{B} are the probabilities of the links.

The solutions \mathbf{D}^* and \mathbf{B}^* to the optimization problem defined in (4) is computed according to (5) [18].

$$\begin{aligned} \mathbf{D}^* &= \left((\mathbf{C}^{(1)})^T \mathbf{C}^{(1)} \right)^{-1} (\mathbf{C}^{(1)})^T \mathbf{S} \mathbf{C}^{(1)} \left((\mathbf{C}^{(1)})^T \mathbf{C}^{(1)} \right)^{-1} \\ \mathbf{B}^* &= \left((\mathbf{C}^{(1)})^T \mathbf{C}^{(1)} \right)^{-1} (\mathbf{C}^{(1)})^T \mathbf{A} \mathbf{C}^{(2)} \left((\mathbf{C}^{(2)})^T \mathbf{C}^{(2)} \right)^{-1} \\ &\text{s.t. } \mathbf{C}^{(1)} \in \{0, 1\}^{|V_1| \times K_1}, \mathbf{C}^{(2)} \in \{0, 1\}^{|V_2| \times K_2}, \\ &\mathbf{D}^* \in R_+^{K_1 \times K_1}, \mathbf{B}^* \in R_+^{K_1 \times K_2} \end{aligned} \quad (5)$$

The complex graph clustering algorithm is described step by step below. For more theoretical details, please refer to [18].

Input: A complex graph $G = (V_1, V_2, E)$, assuming that the number of clusters in V_1 is K_1 and the number of clusters in V_2 is K_2 .

Output: The result of topic clustering $\mathbf{C}^{(1)}$ and the result of image clustering $\mathbf{C}^{(2)}$. Matrix \mathbf{P} where the elements are the link patterns between the clusters in $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$.

1. Given the initial values of $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$, calculate the initial value of \mathbf{D} , \mathbf{B} , and L to generate $L_{\min} = L_{\text{init}}$.
2. Fix \mathbf{D} , \mathbf{B} , and $\mathbf{C}^{(2)}$, then update each element of $\mathbf{C}^{(1)}$ to 1 row by row, and generate the L minimum at each update. Update L_{\min} .
3. Fix \mathbf{D} , \mathbf{B} , and $\mathbf{C}^{(1)}$, then update each element of $\mathbf{C}^{(2)}$ to 1 row by row, and generate the L minimum at each update. Update L_{\min} .
4. Calculate \mathbf{D} and \mathbf{B} using Eq. (5).
5. Repeat steps 2 to 4 until convergence is reached.
6. Calculate the mapping relationship matrix \mathbf{P} between the image clusters and topic clusters according to $P(I|T) = P(I|T')P(T'|T)$, where $P(I|T')$ is the \mathbf{B} matrix and $P(T'|T)$ is the \mathbf{D} matrix.

The complex graph clustering algorithm can cluster the image vertices and topic vertices separately and produce a one-to-one mapping between the topic clusters and the image clusters. During the clustering process, the three relevance relationships will affect each other. During image clustering, images with similar visual contents and close topic contents will form one cluster. During topic clustering, topics with similar visual contents will form one cluster and produce a hot topic.

To make the obtained hot topics understandable to users, keywords are identified in each hot topic. χ^2 is computed to select the words most relevant to a hot topic. Greater values of χ^2 means higher relevance to the hot topic.

4.4 Updating the annotated set

With the hot topics discovered and keywords identified, the annotated image set is then updated to cover wider range of semantics.

Different from the traditional model-based annotation methods, no relevance model or mapping relationships between visual features and semantic concepts are maintained for the annotated set. Hence the computational cost of updating is much

lower and acceptable. To further decrease the cost, images in the set is grouped and organized according to hot topics, dividing the entire set into multiple semantically and visually coherent subsets. Mean feature vectors are computed to represent each subset, whose semantics is expressed with the keywords of the corresponding hot topic.

With the benefit from the data organization, the annotated set can be updated by directly adding new hot topics as new subsets. For instance, considering the situation that many food safety accidents have been discovered, which is not included in the current annotated set, e.g. red yolk stained with Sudan Red, etc. Each accident corresponds to a hot topic which is a set of images and annotations. Then the updated annotated set is the union of the original set and these new hot topic sets. Note that the original subsets remain unchanged.

It is also possible that the update is based on an existing subset of the current annotated set. If new semantics emerges with respect to the topic, then the original topic will split into several sub-topics. And the sub-topics are used to replace the old topic. It can be imagined that at the very beginning, apple only indicates a kind of fruit. So there is only one topic of fruit apple in the annotated set. Then one day the Apple Inc. is founded, and when “apple” is used as a query keyword to search the Internet, two topics will be discovered - fruit apple and Apple Inc. So the old topic of apple in the annotated set is abandoned, and two new topics of fruit apple and Apple Inc. are added.

This update mechanism keeps expanding the semantics covered by the annotated set and eliminating ambiguous topics with the minimal cost. And the form of data organization makes it quite convenient to find the images visually similar to a given one.

5 Search-based image annotation

To annotate a given image, the annotated set will be searched in order to find the images, which are visually similar with the given one. For similarity measurement, please refer to Section 4.1. Then a hypergraph is constructed with the images found and their annotations, on which clustering is performed in order to identify the images semantically relevant or irrelevant to the given one. After filtering out the semantically irrelevant images, keywords are selected from the annotations of the remaining images, which are assigned to the given image.

5.1 Filtering out semantically irrelevant images using a hypergraph

When the set of similar images V_s and the corresponding annotation set T_s are obtained, a hypergraph $G(V_s, T_s)$ can be constructed with images as vertices and annotations as hyperedges, i.e. If two images share the same annotation, then the corresponding two vertices are connected by a hyperedge. Hypergraph G can be represented as a matrix \mathbf{H} as shown in Fig. 2. \mathbf{H} also encodes the co-occurrence relationships between annotations. The graph is then partitioned using spectral clustering [22], forming clusters containing images that have the most similar annotations. Therefore images with different semantics will be well separated, and those that are

semantically irrelevant will form some small clusters, which will be abandoned as analyzed in Section 3.2.

5.2 Annotating the given image

After the semantically irrelevant images are removed, the cluster that is the most visually similar to the given image can be identified, the annotations of which are selected as candidates. Let S denote the identified cluster. Then the final annotations are determined according to the relevance between the given image and candidate annotations in S , as is defined in (6). The conditional probability given the query image $p(t_i|I_q)$ is used to measure the relevance between the query image I_q and the i -th candidate annotation t_i in S . The candidate annotations with high relevance will be preserved and assigned to the given image.

$$\begin{aligned}
 p(t_i|I_q) &= \sum_{I_j \in S} p(t_i|I_j)p(I_j|I_q) \\
 p(t_i|I_j) &= \begin{cases} 1 & \text{if } I_j \text{ is annotated with } t_i \\ 0 & \text{others} \end{cases} \\
 p(I_j|I_q) &\propto \text{similarity}(I_j, I_q)
 \end{aligned} \tag{6}$$

similarity (I_j, I_q) is the visual similarity between image I_j and I_q , which is described in Section 4.1.

6 Experiments and evaluation

We propose a new high-level semantic annotation algorithm based on hot Internet topics. It includes the dynamic update of the training set based on hot Internet topics and search-based image annotation. Through the dynamic update of the training set, the semantics in the training set are more accurate and cover more recent hot Internet topics. For the annotation process, users could get the annotation results of the query image, which contain the keywords of the hot topics. So, in this section, we conduct four groups of experiments to evaluate our overall algorithm.

Experiment 1 Performance evaluation of the search-based image annotation.

Experiment 2 Performance evaluation of the topic discovery and image annotation algorithm.

Experiment 3 Effectiveness evaluation of the dynamic update of the training set based on hot Internet topics.

Experiment 4 Effectiveness evaluation of the annotation algorithm in an Internet-like environment.

6.1 Data set selection and performance measurement

6.1.1 Data set selection

Four data sets are used in our experiments, which cover a wide range of situations.

Dataset1 The NUS-WIDE data set [3] is selected for the experiment 1. It is a web image data set created by NUS's Lab for media search. It contains 269,648 images from Flickr with 425,059 unique tags in total. The data set is divided into two parts: the first part contains 161,789 images for training and the second part contains 107,859 images for testing. In the training set, we remove the tags which occurred less than 50 times. In the meantime, the tags not in WordNet are filtered out via WordNet Stemmer. The final number of the unique tags is 5,018. In the testing set, the number of manually labeled tags per image varies from 2 to more than 100, with an average number of about 30. The WordNet stemmer is also used to do stemming, and then we remain the tags which are occurred more than 50 times in the testing set. Finally, the average number of manually labeled tags per image is about 10. This data set also has manual annotations as the ground truth with 81 concepts in total, which belong to different categories. The ground truth for the testing set is the processed manual annotations and related concepts.

Dataset2 This is a hot topic corpus, which is built from the searching results (include web pages and related images) on the Internet for 34 topics. These topics include below:

1. 10 concepts which get the lowest accuracy rates in the experiment 1: In the experiment 1, we calculate the annotation performance of every concept. And then 10 concepts with the lowest average accuracy rates in the 81 concepts are selected. They are: C1: Earthquake, C2: Statue, C3: Rainbow, C4: Running, C5: Wedding, C6: Book, C7: Castle, C8: Flags, C9: Temple, C10: Train.
2. 20 food safety accidents during 2008 to 2012 in China: We collected the food safety accidents reported on the Internet and organized them into 20 accidents, i.e., E1: Sanlu milk powder, E2: Paraffin chafing-dish material, E3: Poisoned capsules, E4: Jinhao tea oil, E5: McDonald's chicken, E6: Plasticizer event, E7: Poisoned yoghurt, E8: Sudan Red accident, E9: Trench oil, E10: Small lobster event, E11: Lean meat powder event, E12: Poisoned bread, E13: Maggots-orange, E14: Burst watermelon, E15: Toxic bird's nest, E16: Turbot accident, E17: Poison bean sprouts, E18: Maggots-sausage, E19: Poisoned ginger event, E20: Deteriorating rice event.
3. 4 visual polysemous words: We evaluate the effectiveness and performance of complex graph clustering by using four visual polysemous words as query keywords, i.e., "apple", "tiger", "mouse" and "shark".

The keywords from these 34 topics are used to perform the searching on the Internet and the results of web pages and images are downloaded. Removing the duplicated ones, there are 450–1300 text pages and 300–800 images collected for each topic. In the results, about 150–200 images are selected as the testing images in the experiment 2 and the remaining images are used as the training set.

Dataset3 This is an updated data set. In the experiment 3, we use the Dataset2 to update the Dataset1 and the result is the Dataset3. We will do the annotation experiment 1 again on this data set and compare with the results on Dataset1.

Dataset4 This data set is constructed by updating the training set in Dataset1 to form 196 subtopics based on the original 81 concepts. And 100 images are selected randomly from each subtopic, resulting in a training set of 19,600 images. The test set

of 107,859 images remains unchanged, so the ratio of training and test set size is 1:5.5. Furthermore, in order to compare the performances with different ratios, we also used the entire set (1:0.6) and select 10 images from each subtopic (1:55).

6.1.2 Performance measurement

The average precision rate (Av_P) and the average recall rate (Av_R) as defined in (7) are used to measure the performance of our image annotation algorithm.

$$P(I_i) = \frac{|A_i \cap G_i|}{|A_i|}$$

$$R(I_i) = \frac{|A_i \cap G_i|}{|G_i|} \quad (7)$$

$$Av_P = \frac{1}{N} \sum_{i=1}^N P(I_i)$$

$$Av_R = \frac{1}{N} \sum_{i=1}^N R(I_i) \quad (8)$$

N denotes the total number of images. A_i denotes the set of result annotations assigned to the image by our algorithm and G_i denotes the ground truth for image I_i .

Furthermore, coverage rate (Cov_rate), as defined in (9), is used to evaluate the number of keywords in the vocabulary that are used to annotate the query images.

$$Cov_rate = \frac{|\bigcup_{i=1}^N A_i|}{|V|} \quad (9)$$

where V is the annotation vocabulary of the evaluation data set.

6.2 Experiments on Dataset1

We evaluate the search-based image annotation task on the Dataset1 and compare the results with SBIA [32] and LTA [34]. The detailed configuration of the experiments is described as below.

- **SBIA:** It requires both the query image and an initial correct query keyword as the input. So, we assign the 81 concepts in the Dataset1 to the corresponding query images as the initial query keywords.
- **LTA:** For one query image to be annotated, it also requires several initial keywords. The same as SBIA, we assign the 81 concepts in the Dataset1 to the corresponding query images as the first keyword. The other 2 initial keywords will be selected from their manual annotations sequentially.
- **Our algorithm:** we require that the training set is organized by the hot topics. However, since the images in the Dataset1 are coming from NUS-WIDE, they have not been organized according to the hot topics. So, we need to leverage the 81 concepts to organize the training set. We calculate the color histogram of each image as its visual feature and get the mean value for each concept on this feature. This mean value will be used as the cluster center for this concept. Each

concept is indexed by a cluster center. Then for a query image, we find several neighbor concepts whose centers are closest to the query image, after which similar images are identified among those belonging to the selected concepts. With these images and their tags constructing a hypergraph, clustering is performed in order to annotate the query image.

- The number of output tags is set to 10 in the experiments.

Table 1 shows the annotation performance of three algorithms. From the results, our algorithm is clearly better than SBIA and LTA on annotation performance. Both SBIA and LTA require one or several correct initial keywords to start with, this helps them greatly on the quality of the final annotations. Even though, our algorithm does not need any correct initial keyword but still achieves better performance. Although SBIA is similar with our algorithm as the search-based mechanism and leverages the visual similarity to search for the candidates, it lacks of the filtering process based on hypergraph clustering. We don't pick the keywords directly based the annotations of all the candidates. Instead, we leverage the semantic relevance relationship of the images and annotation words to filter out the images with irrelevant semantics. Through hypergraph clustering, several clusters of candidates are formed. And then the cluster of the candidates with the most similar visual features to the query image will be selected and the final annotation words will be picked from this cluster. So the final annotations are much more accurate. For LTA, it considered three relevance relationships: tag-to-tag co-occurrence relationship based on the text characteristics (TC), tag content correlation based on visual similarity (TCC) and image conditioned tag correlation (ITC). For example, when calculating the TCC for the tag pair (t_i, t_j) , LTA will collect all the images with the annotation of t_i and t_j respectively and then compute the visual similarity of these two image sets based on VLM model. The visual similarity of these two image sets represents the visual similarity of the two tags. However, because of the well-known semantic gap problem, even the semantics of t_i and t_j are the same, their visual representations may be quite different. So, building the tag content correlation based on the images with these tags suffers from this semantic gap problem. We could achieve better performance because a hypergraph clustering mechanism in our algorithm could filter out the images with irrelevant semantics. The candidate cluster will have better consistence in terms of visual features and semantics. So, the final annotation results will be more consistent and accurate.

From Table 1, we can also see that our algorithm has better coverage rate than SBIA and LTA. This is because SBIA and LTA are all tending to select the most common keywords for the testing images. In our algorithm, we calculate the annotation probability for each keyword in the candidate cluster. The annotation probability is related with the similarity between the query image and the candidate image. So, even the keywords that are used less often in the cluster can be selected as the keywords to annotate the images.

Table 1 Results of the comparison

	Our algorithm	SBIA [32]	LTA [34]
Precision	0.23	0.19	0.17
Recall	0.17	0.15	0.14
Con_Avg	0.24	0.21	0.18

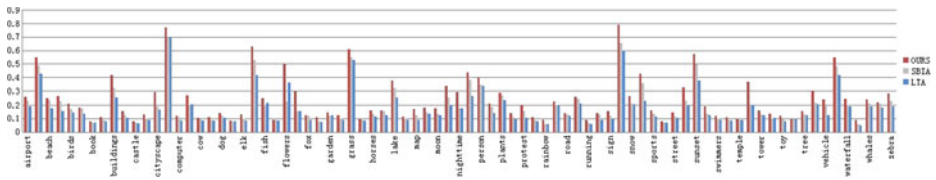


Fig. 3 Average precisions of the 81 concepts with the 3 approaches

Figure 3 shows the distribution of the average accuracy rates on 81 concepts for these 3 algorithms. From the distribution curve, these 10 concepts have the lowest accuracy rates in our algorithm: “earthquake”, “statue”, “rainbow”, “running”, “wedding”, “book”, “castle”, “flags”, “temple”, “train”. After the analysis, we found three reasons for these concepts which lead to the low accuracy rates. First, in the training set, the number of images in these 10 concepts is lower than other concepts. So, it is hard to have positive effects according to the query image in these concepts. Second, there are many noisy tags existed in these concepts, such as in the concepts of “earthquake”, “book”, “castle”, and “running”, etc. Third, the semantic gap problem also has impact on some concepts and reduces the annotation quality. For example, in the “rainbow” concept, there are two types of topics in the images: rainbow-nature and rainbow-band. These two topics are all annotated with “rainbow”, but have huge difference in visual characteristics. So, the learning model is not effectively built up for this concept. In the experiment 2, we will update the training set for these 10 concepts based on hot Internet topics and evaluate the improvement on the annotation for the training set in experiment 3.

6.3 Experiments on Dataset2

6.3.1 The parameters in experiment 2

As described in Section 6.1.1, we search the Internet for web pages and related images according to 34 topics. The search results will form the hot topic corpus. For each topic, we leverage LDA algorithm to build the topic model for the web pages. Before this experiment 2, we set the number of LDA topics to 5, 8, 10, 15, 20, 30, 40, and 50 and evaluate the results of LDA training and complex graph clustering in order to find the best topic number for LDA algorithm and clustering number for complex graph clustering. After the experiments, we found that the clustering result is the best when the topic number for LDA is set to 10 and the cluster number is set to 3. So, in the experiment 2, we always set the topic dimension for each web page to 10 and the number of clusters to 3. At the same time, the number of keywords for each hot topic is set to 10 consistently.

6.3.2 Measure the effectiveness of three relationships

We leverage three relevance relationships in our algorithm for abstract semantic modeling: topic-to-topic co-occurrence relationship, topic-to-image relevance relationship and image-to-image similarity relationship. Through the complex graph clustering, the image set will be clustered into several clusters with different semantics and visual features. In order to evaluate the effectiveness of these three

relationships, we conduct three groups of experiments to evaluate the performance of clustering based on different relationships. The Normalized Mutual Information (NMI) is used as the quantitative measurement for clustering performance evaluation.

1. Based on the three relevance relationships proposed in this paper, the complex graph clustering is performed and the NMI is calculated according to the image clustering results.
2. Based on two relevance relationships (topic-to-topic co-occurrence and topic-to-image relevance), the complex graph clustering is performed and the NMI is calculated according to the clustering result.
3. The baseline for this evaluation is the NMI result of K-means algorithm. The LDA topics for the images are used as the document features. Together with the visual features of the images, we use K-means to do the clustering as the baseline.

NMI is the standard performance evaluation method used for clustering. Let k be the number of clusters and let $\lambda = (\lambda_1, \dots, \lambda_N)$ be the cluster vector where $\lambda_i = 1, \dots, k$. $\lambda_i = j$ denotes that the i -th item belongs to the cluster C_j . If $\lambda^{(a)}$ and $\lambda^{(b)}$ are clustering result and the ground truth vectors respectively, the NMI criterion Φ can be calculated as follows [27]:

$$\Phi^{(NMI)}(\lambda^{(a)}, \lambda^{(b)}) = \frac{\sum_{h=1}^N \sum_{l=1}^N n_{hl} \log \left(\frac{n_{hl}}{n_h^{(a)} \cdot n_l^{(b)}} \right)}{\sqrt{\left(\sum_{h=1}^N n_h^{(a)} \log \frac{n_h^{(a)}}{n} \right) \left(\sum_{l=1}^N n_l^{(b)} \log \frac{n_l^{(b)}}{n} \right)}} \tag{10}$$

where $n_h^{(a)}$ is the number of items in cluster C_h in $\lambda^{(a)}$, while $n_l^{(b)}$ denotes the number of items in cluster C_l in $\lambda^{(b)}$. n_{hl} is the number of items in both cluster C_h and cluster C_l . Based on this definition, the clustering result is better if $\Phi^{(NMI)}(\lambda^{(a)}, \lambda^{(b)})$ is bigger. The theoretical maximum is 1 for $\Phi^{(NMI)}(\lambda^{(a)}, \lambda^{(b)})$.

6.3.3 Performance evaluation of semantic annotation based on hot topics

(1) Evaluation of the annotation results for 10 concepts

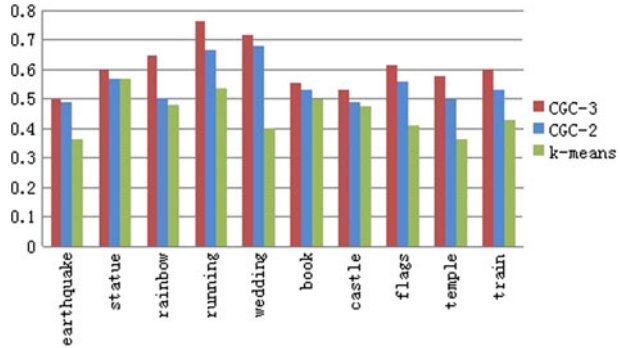
As in Table 2, after the complex graph clustering, there are several sub-topics formed for each concept.

From Table 2, there are several keywords in the concepts which are visually polysemous and semantically ambiguous. It means one concept may have multiple

Table 2 Some of the concepts and the corresponding sub-topics

Original concepts	Sub-topics
Earthquake	Earthquake damage, rescue
Statue	Statue
Rainbow	Rainbow, rainbow band
Running	Runner, shoes
Wedding	Wedding flower, wedding ring, wedding
Book	Book
Castle	Castle, TV show
Flags	Flags
Temple	Temple, Buddha
Train	Train

Fig. 4 NMI of clustering with respect to the 10 concepts



visual representatives. In the training set, although several images may belong to the same concept, their visual characteristics could be quite different. So, it will be very challenging to learn the correct relevance relationship based on this kind of training set. In our algorithm, we use the complex graph clustering technology to category the images in one concept into sub-topics. The images in each sub-topic will have almost consistent visual characteristics and semantics.

The NMI data for clustering results based on different relevance relationships is showed in Fig. 4. As the figure shown, the clustering based on three relationships delivered the best performance and it can separate the visual polysemous or ambiguous concepts into several sub-topics. By only using the topic-to-topic co-occurrence relationship and topic-to-image relevance relationship, the visual similarity between the images is ignored. So, even though the clustering result may have similar semantics, the visual difference in one cluster may still be very significant.

Figure 5 shows examples of the hot topics and corresponding keywords, such as rainbow, and running.

(2) *Evaluation of the annotation results for visual polysemous words*

We evaluated the effectiveness and performance of three relevance relationships using four visual polysemous words as initial query keywords, i.e., “apple”, “tiger”,

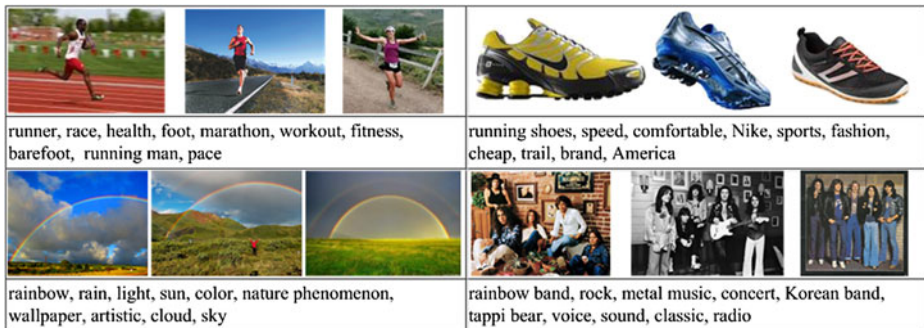


Fig. 5 Examples of sub-topics and the corresponding keywords





	
<p>tiger, forest, stripes, animals, Manas, jungle, Panthera, Bengal, cats, Bhutan</p>	<p>Tiger Woods, golf, golfer, PGA TOUR, championship trophy, scandal, divorce</p>
	
<p>house mouse, tail, mice, rodents, pestilence, carrier, disease, hole, mousetrap,</p>	<p>wireless optical mouse, computers, pointing device, sensor, scroll wheel, buttons</p>

Fig. 6 Examples for polysemous words “tiger” and “mouse”

“mouse” and “shark”. The experimental results are shown in Figs. 6 and 7. For the “apple” keyword, the clustering results contained three sub-topics: apple as a fruit, Apple product and the Apple Inc. There were two sub-topics for “tiger”: tiger as an animal and Tiger Woods. There were two sub-topics for the “mouse” keyword: mouse as an animal and a computer mouse. There were three sub-topics for the “shark” keyword: the shark as an animal, Shaq O’Neal, and the band Shark. Thus, the input query keywords were semantically ambiguous but we could determine the correct clustering results based on the visual content. We also annotated each cluster correctly according to their semantics. The NMI results for clustering are shown in Fig. 8. The complex graph clustering method considered three types of relationships, so it delivered a better performance than complex clustering for two types of relationships and the K-means algorithm for a single relationship.

(3) *Evaluation of the image annotation performance for hot Internet topics*

In this experiment, we collected the food safety accidents reported on the Internet and organized them into 20 categories as described in Section 6.1.1. NMI of clustering results with respect to the food safety accidents are shown in Fig. 9, which demonstrates that the image clusters and topic clusters were consistent with the visual contents and semantics. The annotation results extracted from the topic clusters conveyed the semantics of the food safety accidents correctly.

For example, Fig. 10 shows that for image “capsule”, the extended annotation included Chromium, heavy metal, and poisonousness. For image “yoghourt”, the

		
<p>apple, red, tasty, Vitamins, minerals, nutrient, pectin, fiber, lose weight</p>	<p>Steve Jobs, Apple, CEO, address, iPhone, geek, resign, health issue, tumor</p>	<p>mac, apple, smartphone, ipad, ipod, PC, jobs, refurbished, white</p>
		
<p>O’Neal, Shaquille, shark, NBA, Lakers, midfielder.championship,Miami Heat, retirement</p>	<p>Big Shark, rock band, music festival, album, Houhai, Modern Sky, concert, star, club</p>	<p>great white shark, teeth, JAWS, killer, predator, fin, prey, protection, ocean</p>

Fig. 7 Examples for polysemous word “apple” and “shark”

Fig. 8 NMI of clustering with respect to the polysemous words

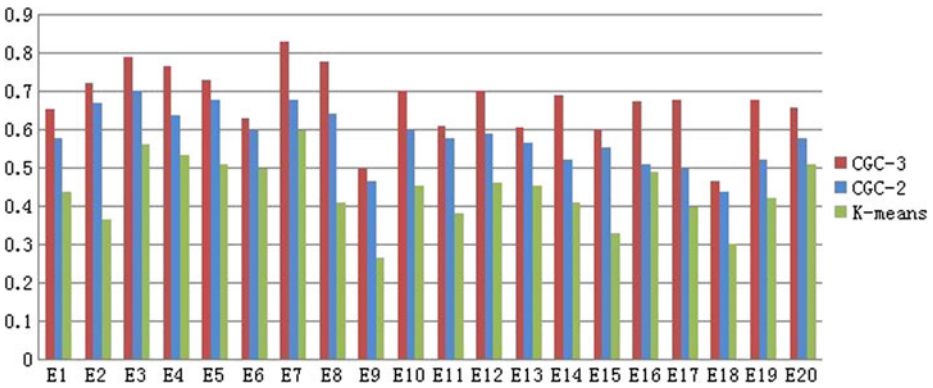
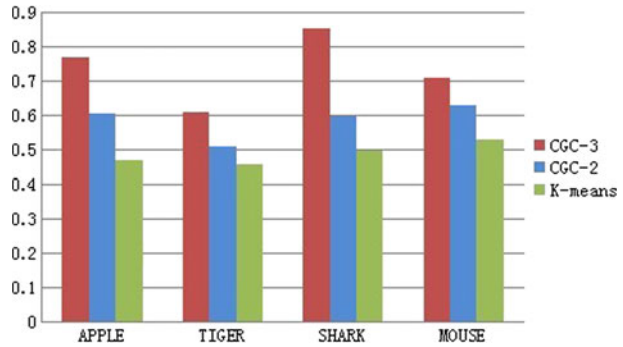


Fig. 9 NMI of clustering with respect to food safety accidents










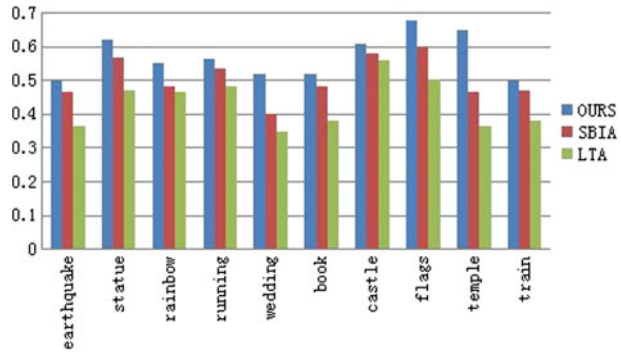
			capsule, Chromium, heavy metal, poisonousness, Melamine, industrial waste, Pharmaceutical, Pharmacies, Quality, Biological Drug
			yoghurt, nutrition, additive, taste, protein, gelatin, agar, raw material, jelly, Mengniu Dairy
			red yolk, duck egg, Sudan, Sudan Red, Baiyangdian, color, Red Medicine, carcinogenic, nutritive value, additive

Fig. 10 Examples for food safety accidents

Fig. 11 Precision of the 10 topics



annotation was “gelatin” because this food safety issue was related to the illegal addition of gelatin to yoghurt. The annotations of the images for “red, yolk, and duck egg” had the abstract semantics “Sudan red”, “additive”, etc.

6.4 Experiments on Dataset3

In the experiment 2, 47 keywords as hot topics are constructed from 34 topics. Now, we add the new hot topics into the Dataset1 and replace the original 10 concepts with lowest accuracy rates. The update dataset is called the Dataset3. We will perform the following experiments in this section.

1. Annotation experiments on the original images in these 10 concepts.
2. Annotation experiments based on 20 hot topics on food safety accidents.

As Fig. 11 shown, based on the updated training set for these 10 concepts, the annotation results of SBIA, LTA and our algorithm are all improved. As for our algorithm, the average precision of the 10 concepts on the updated annotated set are improved by 5.4x. The reasons are two follows. After updating, the concepts

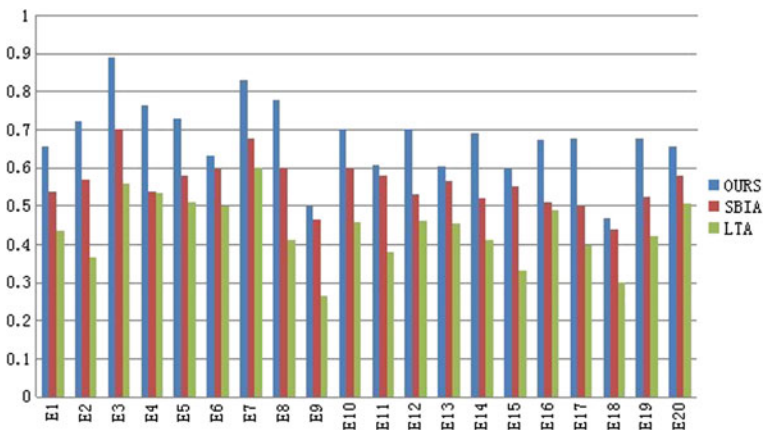


Fig. 12 Precision of 20 topics of food safety accidents

with polysemous or ambiguous images are separated into several sub-topics. In each sub-topic, the image set and the text set have better consistent semantics and visual features. In the meanwhile, the keywords based on hot topics have been merged into the annotations of related images so that the annotation accuracy in the training set gets improved. So, the final annotation performance is improved as well. Our experiments demonstrate that an effective update mechanism for the training set is highly desirable for image annotation.

Figure 12 shows the annotation performance of 20 topics on food safety accidents. The average precision of all 20 topics with our approach, SBIA and LTA are 0.67, 0.55 and 0.44 respectively. The results demonstrate the effectiveness of the update and the superiority of our approach. Figure 13 shows some examples of our annotation results.

6.5 Experiments on Dataset4

Most previous annotation algorithms construct annotation mapping models based on one annotated training set, and then test images will be annotated according to these mapping models. In their experiments, the number of images in the training set is generally larger than that in the test set. This configuration does not match the real environment since the number of annotated images is much smaller than that of un-annotated ones on the Internet. So in this experiment, the ratio of the image numbers in the training and test set is made to be 1:5.5 in order to evaluate the annotation algorithm in an environment closer to the real one.

As shown in experiment 1, in the search-based annotation sub-task, we first leveraged the visual features, such as color histogram, and wavelet texture, to find out the candidates from the training set, which are similar to the query image. Then the keywords were extracted from this candidate set for annotation. Because of the semantic gap problem, the images with similar visual features may have different semantics. Hypergraph clustering is used to remove the non-consistent images from the candidate set in order to mitigate the semantic gap problem. So, based on the




Test Image	Human Annotation	Annotation by Our Algorithm
	watermelon explosion, swelling agent, farmer, sweet, chemical agent	watermelon, watermelon explosion, fruit, boost, growth, chemical free, food safety, harmful, fertiliser, forchlorfenuron
	earthquake, pakistan, azadkashmir, muzaffarabad	earthquake damage, magnitude, seismic waves, energy, destroyed building, economic cost, landslide, tsunami, rubble, ruins
	orange, fish, ilovenature, underwater, nemo, scuba, diving, clownfish, anemone, perhentian, specanimal	fish, nemo, anemone, clownfish, underwater, orange, sea, diving, scuba, coral

Fig. 13 Examples of our annotation results

Table 3 Results with different combination of features

Features	CH+WT	CH+SIFT (%)	SIFT+WT (%)	SIFT+SIFT (%)
Mean precision	56.8	51.9	59.5	57.7
Mean recall	61.7	56.3	64.4	62.1

filtering mechanism, the performance of our algorithm would not be so correlated with the visual features exploited. In experiment 4, the annotation is based on the following four combinations of the features.

- Color histogram and wavelet texture
- Color histogram and SIFT
- SIFT and wavelet texture
- SIFT and SIFT

The results in Table 3 show that the annotation performances based on different combinations of features are quite close. This proves the effectiveness of our annotation algorithm that the candidates after filtering have similar visual features and semantics. And moreover, compared to that in the test set, the number of images in the training set is very small, which is similar to the real running environment on the Internet. In such an environment, our algorithm can still achieve good results.

In addition, with the combination of color histogram and wavelet texture, we also performed experiments on the whole training set of 161,789 images, as well as a subset of 1,960 images, which is obtained by selecting 10 images from each subtopic. Precisions with training sets of size 161,789, 19,600 and 1,960 are 60.2, 56.8 and 36.0 % respectively. Note that when the training set size decreases to approximately 10 % of the original one, the precision reduces by 5.6 and 36.6 % respectively. The 36.6 % decrease is mainly due to the fact that the annotations covered by the training set (1,960 images) are not sufficient to annotate all the test images. And using the entire set (161,789 images) provides only minor improvements over the training set of 19,600 images since the additional images in the entire set actually share the same annotations with others. It indicates that the annotation coverage of the training set is more important than its size.

7 Conclusion

In this study, we develop a new high-level semantic annotation method for images based on hot Internet topics. This method has two sub tasks: search-based image annotation and the dynamic update of the training set based on hot Internet topics. This method exploits the large-scale image resources available on the Internet for image annotation and regularly updates the training set from hot Internet topics in an efficient way. We propose a new method to model the abstract semantics for images. Three sets of relationships between topics and images are exploited. And through the complex graph clustering, the hot Internet topics are extracted from images with similar visual contents. From the experiments, we have demonstrated the effectiveness of three relationships and the complex graph clustering, which make sure that images with similar visual contents and close topics will form one cluster. The keywords from this cluster are good representatives for the hot topics. The

dynamic update mechanism of the training set addresses the issue of the huge computing cost in traditional update methods, which require to re-calculate the whole relevance relationships between tags and visual features of images. The experiments also show that the updated training set can deliver better annotation results since it can reduce the impact of the semantic gap problem for visual polysemous words. The search-based image annotation can effectively filter out the semantic irrelevant images via hypergraph mechanism. We calculate the annotation probability for each keyword in the candidate cluster. The annotation probability is related with the similarity between the query image and the candidate images. So, even the tags that are used less often in the cluster can be selected as the keywords to annotate the images. The experiments show that its annotation performance is better than the state-of-the-art algorithms.

The future work of our research is to look at its feasibility on large-scale data center and evaluate the computing requirement in the Cloud.

Acknowledgements This research study was supported by the National Basic Research Program of China (973 Program) (2012CB821206), the National Natural Science Foundation of China (No. 91024001, No. 61070142), the Beijing Natural Science Foundation (No. 4111002), and the Fundamental Research Funds for the Central Universities (No. 2013RC0306).

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

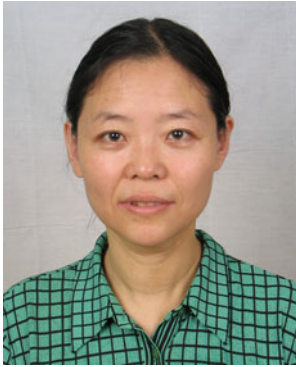
1. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
2. Chang E, Goh K, Sychay G, Wu G (2003) Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Trans Circuits Syst Video Technol* 13(1):26–38. doi:[10.1109/TCSVT.2002.808079](https://doi.org/10.1109/TCSVT.2002.808079)
3. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval, CIVR '09. ACM, New York, pp 48:1–48:9. doi:[10.1145/1646396.1646452](https://doi.org/10.1145/1646396.1646452)
4. Cusano C, Ciocca G, Schettini R (2003) Image annotation using svm, pp 330–338. doi:[10.1117/12.526746](https://doi.org/10.1117/12.526746)
5. Dai W, Chen Y, Xue GR, Yang Q, Yu Y (2009) Translated learning: transfer learning across different feature spaces. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) *Advances in neural information processing systems*, 21, pp 353–360
6. Deschacht K, Moens MF (2007) Text analysis for automatic image annotation. In: Proceedings of the 45th annual meeting of the association of computational linguistics. Association for Computational Linguistics, Prague, pp 1000–1007
7. Duygulu P, Barnard K, Freitas J, Forsyth D (2006) Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Heyden A, Sparr G, Nielsen M, Johansen P (eds) *Computer vision ECCV 2002. Lecture notes in computer science*, vol 2353. Springer Berlin Heidelberg, pp 97–112
8. Eakins JP (1996) Automatic image content retrieval—are we getting anywhere? De Montfort University, Milton Keynes, pp 123–135
9. Feng S, Manmatha R, Lavrenko V (2004) Multiple bernoulli relevance models for image and video annotation. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer vision and pattern recognition, 2004. CVPR 2004, vol 2, pp II–1002–II–1009. doi:[10.1109/CVPR.2004.1315274](https://doi.org/10.1109/CVPR.2004.1315274)
10. Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proc Natl Acad Sci USA* 101(Suppl 1):5228–5235. doi:[10.1073/pnas.0307752101](https://doi.org/10.1073/pnas.0307752101)

11. Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval, ACM, New York, SIGIR '03, pp 119–126. doi:[10.1145/860435.860459](https://doi.org/10.1145/860435.860459)
12. Kang F, Jin R (2005) Symmetric statistical translation models for automatic image annotation. In: Proceedings of the 2005 SIAM international conference on data mining, pp 21–23
13. Kennedy LS, Chang SF, Kozintsev IV (2006) To search or to label?: predicting the performance of search-based automatic image classifiers. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval, MIR '06. ACM, New York, pp 249–258. doi:[10.1145/1178677.1178712](https://doi.org/10.1145/1178677.1178712)
14. Lavrenko V, Manmatha R, Jeon J (2004) A model for learning the semantics of pictures. In: Thrun S, Saul L, Schölkopf B (eds) Advances in neural information processing systems 16. MIT Press, Cambridge, MA, pp 553–560
15. Li J, Wang J (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans Pattern Anal Mach Intell* 25(9):1075–1088
16. Li X, Snoek CG, Worring M (2008) Learning tag relevance by neighbor voting for social image retrieval. In: Proceedings of the 1st ACM international conference on multimedia information retrieval, MIR '08. ACM, New York, pp 180–187. doi:[10.1145/1460096.1460126](https://doi.org/10.1145/1460096.1460126)
17. Liu J, Wang B, Li M, Li Z, Ma W, Lu H, Ma S (2007) Dual cross-media relevance model for image annotation. In: Proceedings of the 15th international conference on multimedia, MULTIMEDIA '07. ACM, New York, pp 05–614. doi:[10.1145/1291233.1291380](https://doi.org/10.1145/1291233.1291380)
18. Long B, Zhang M, Yu PS, Xu T (2008) Clustering on complex graphs. In: Proceedings of the 23rd national conference on artificial intelligence, vol 2, AAAI'08. AAAI Press, pp 659–664
19. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* 60(2):91–110
20. Manjunath B, Ma W (1996) Texture features for browsing and retrieval of image data. *IEEE Trans Pattern Anal Mach Intell* 18(8):837–842. doi:[10.1109/34.531803](https://doi.org/10.1109/34.531803)
21. Monay F, Gatica-Perez D (2007) Modeling semantic aspects for cross-media image indexing. *IEEE Trans Pattern Anal Mach Intell* 29(10):1802–1817. doi:[10.1109/TPAMI.2007.1097](https://doi.org/10.1109/TPAMI.2007.1097)
22. Ng AY, Jordan MI, Weiss Y (2001) On spectral clustering: analysis and an algorithm. *Adv Neural Inf Process Syst* 14:849–856
23. Qi GJ, Hua XS, Rui Y, Tang J, Mei T, Zhang HJ (2007) Correlative multi-label video annotation. In: Proceedings of the 15th international conference on multimedia, MULTIMEDIA '07. ACM, New York, pp 17–26. doi:[10.1145/1291233.1291245](https://doi.org/10.1145/1291233.1291245)
24. Rui X, Li M, Li Z, Ma WY, Yu N (2007) Bipartite graph reinforcement model for web image annotation. In: Proceedings of the 15th international conference on multimedia, MULTIMEDIA '07. ACM, New York, pp 585–594. doi:[10.1145/1291233.1291378](https://doi.org/10.1145/1291233.1291378)
25. Saenko K, Darrell T (2009) Unsupervised learning of visual sense models for polysemous words. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) *Adv Neural Inf Process Syst* 21:1393–1400
26. Shapiro LG, Stockman GC (2003) *Computer vision*. Prentice Hall, Englewood Cliffs, NJ
27. Strehl A, Ghosh J (2003) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* 3:583–617. doi:[10.1162/153244303321897735](https://doi.org/10.1162/153244303321897735)
28. Tsirikla T, Diou C, Vries A, Delopoulos A (2011) Reliability and effectiveness of click-through data for automatic image annotation. *Multimed Tools Appl* 55(1):27–52. doi:[10.1007/s11042-010-0584-1](https://doi.org/10.1007/s11042-010-0584-1)
29. Wang C, Jing F, Zhang L, Zhang HJ (2006) Scalable search-based image annotation of personal images. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval, MIR '06. ACM, New York, pp 269–278. doi:[10.1145/1178677.1178714](https://doi.org/10.1145/1178677.1178714)
30. Wang J, Geman D, Luo J, Gray R (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](https://doi.org/10.1109/TPAMI.2008.231)
31. Wang XJ, Zhang L, Jing F, Ma WY (2006) Annosearch: image auto-annotation by search. In: 2006 IEEE Computer Society conference on computer vision and pattern recognition, vol 2, pp 1483–1490. doi:[10.1109/CVPR.2006.58](https://doi.org/10.1109/CVPR.2006.58)
32. Wang XJ, Zhang L, Li X, Ma WY (2008) Annotating images by mining image search results. *PIEEE Trans Pattern Anal Mach Intell* 30(11):1919–1932. doi:[10.1109/TPAMI.2008.127](https://doi.org/10.1109/TPAMI.2008.127)
33. Wu F, Han YH, Zhuang YT (2010) Multiple hypergraph clustering of web images by mining word2image correlations. *J Comput Sci Technol* 25(4):750–760. doi:[10.1007/s11390-010-1058-7](https://doi.org/10.1007/s11390-010-1058-7)

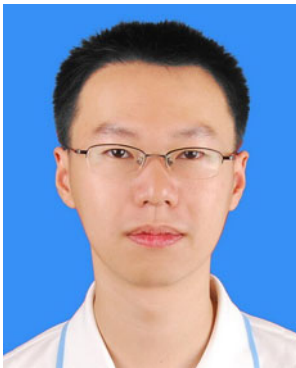
34. Wu L, Yang L, Yu N, Hua XS (2009) Learning to tag. In: Proceedings of the 18th international conference on world wide web, WWW '09. ACM, New York, pp 361–370. doi:[10.1145/1526709.1526758](https://doi.org/10.1145/1526709.1526758)
35. Xia D, Wu F, Zhuang Y (2008) Search-based automatic web image annotation using latent visual and semantic analysis. In: Huang YM, Xu C, Cheng KS, Yang JF, Swamy M, Li S, Ding JW (eds) Advances in multimedia information processing—PCM 2008. Lecture Notes in Computer Science, vol 5353. Springer Berlin Heidelberg, pp 842–845
36. Xiang Y, Zhou X, Chua TS, Ngo CW (2009) A revisit of generative model for automatic image annotation using markov random fields. In: IEEE Conference on computer vision and pattern recognition, 2009. CVPR 2009, pp 1153–1160. doi:[10.1109/CVPR.2009.5206518](https://doi.org/10.1109/CVPR.2009.5206518)
37. Xiang Y, Zhou X, Liu Z, Chua TS, Ngo CW (2010) Semantic context modeling with maximal margin conditional random fields for automatic image annotation. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR), pp 3368–3375. doi:[10.1109/CVPR.2010.5540015](https://doi.org/10.1109/CVPR.2010.5540015)
38. Zhang X, Li Z, Chao W (2013) Improving image tags by exploiting web search results. *Multimed Tools Appl* 62(3):601–631. doi:[10.1007/s11042-011-0863-5](https://doi.org/10.1007/s11042-011-0863-5)
39. Zhu X, Goldberg AB, Eldawy M, Dyer CR, Strock B (2007) A text-to-picture synthesis system for augmenting communication. In: Proceedings of the 22nd national conference on artificial intelligence, vol 2, AAAI'07. AAAI Press, pp 1590–1595



Xiaoru Wang received her B.S. and M.S. degrees in computer science and technology from Beijing University of Posts and Telecommunications, China, in 1997, 2001 respectively. Now she is the doctoral student in Beijing University of Posts and Telecommunications. Her major interests are multimedia retrieval, image tagging and data mining.



Junping Du was born in 1963. Now, she is a full professor and Ph.D. tutor with the School of Computer Science and Technology, Beijing University of Posts and Telecommunications. Prof. Du is the Director of the Computer Applications Center in Beijing University of Posts and Telecommunications University. Her research interests include artificial intelligence and intelligent information system.



Shuzhe Wu is currently a master student at University of Chinese Academy of Sciences. He received his BEng degree in computer science and technology from Beijing University of Posts and Telecommunications in 2013. His research interests include image segmentation, annotation and understanding.



Xu Li is currently looking for his master degree at Beijing University of Posts and Telecommunications. He received his bachelor degree in computer science and technology from Beijing University of Posts and Telecommunications in 2013. His primary research focus is in the area of computer vision. He also works on data mining and cloud computing.



Haiming Xin received his BEng degree in computer science and technology from Beijing University of Posts and Telecommunications in 2013. His research interests include semantic annotation and image feature extraction.



Yu Zhang is currently a master student of Beijing University of Posts and Telecommunications. She received her bachelor degree in mathematics and applied mathematics from Ocean University of China in 2011. Her research interests include image feature extraction, recommended algorithm.



Fu Li received his B.S. and M.S. degrees in physics from Sichuan University, China, in 1982 and 1985, respectively, and his Ph.D. degree in electrical engineering from the University of Rhode Island in 1990. Since 1990, he has been with Portland State University where he is currently a Full Professor of Electrical and Computer Engineering. His research interests include signal, image, and video processing, as well as wireless, network, and multimedia communications.