Neurocomputing 119 (2013) 41–48

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Hybrid image summarization by hypergraph partition

Minxian Li [a,*], Chunxia Zhao [a], Jinhui Tang [a,b]

[a] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, PR China
[b] State Key Lab. for Novel Software Technology, Nanjing University, Nanjing, PR China

ARTICLE INFO

*Article history:*
Received 29 October 2011
Received in revised form
24 January 2012
Accepted 7 February 2012
Communicated by: Chennai Guest Editor
Available online 3 January 2013

*Keywords:*
Hybrid image summarization
Hypergraph
Hypergraph partition

ABSTRACT

The objective of hybrid image summarization is selecting a few visual exemplars and semantic exemplars of a large-scale image collection and organizing them to represent the collection. In this paper, we present a framework for hybrid image summarization in which social images and corresponding textual information are taken as vertices in a hypergraph and the task of image summarization is formulated as the problem of hypergraph partition. A generalized spectral clustering technique is adopted to solve the hypergraph partition problem. Besides, we design two representativeness score functions to select the visual exemplars and semantic exemplars. The main advantages of the proposed approach are two-fold: (1) the hypergraph framework takes advantage of homogeneous correlations within images and tags, respectively, as well as heterogeneous relations between them, this characteristic enhances the summarization performance; and (2) we take both visual and semantic representativeness into count to select exemplars, so that the image-tag exemplars are more representative for each cluster. The experimental comparisons to the other method are conducted on some common queries for a real internet image collection. User-based evaluation demonstrates the effectiveness of the proposed approach.



## 1. Introduction

Rapid advances of cameras and Web 2.0 technology have resulted in the proliferation of image data in the internet. It is reported that there are over 3 million photos being uploaded to Flickr every day. However, image data are usually unorganized on existing photo-sharing web sites, so that it makes finding desired photos and quick overview of an image collection quite difficult. For example, Flickr [1] presents an overview of an image collection by showing the top images, which lists in seemingly random order, and Picasa [2] presents an image collection by allowing consumers to select images manually, which is not convenient for consumers particularly in a large number of images.

Recently, many approaches have been proposed to annotate images for semantic-based image search [3,4]. However, the image search results still have lots of images, which are difficult for users to explore. In order to better organize and browse large-scale image collections, image summarization techniques have received extensive attention recently. Despite intense research efforts [5–14], the results of the existing image summarization techniques are still not satisfactory enough.

Some of them [5–7] only use textual information or geotags. Clough et al. [5] used textual caption data and the concept of subsumption to construct a hierarchy of images. Schmitz [6] used a similar method but relied on Flickr tags. Jaffe et al. [7] summarized a set of images by using only tags and geotags. None of these approaches exploits the visual information.

Some of them [8–11] are primarily based on the low-level features of images, rarely considering the high-level semantics in the images, although most of the internet images have rich text information describing their semantics. Simon et al. [8] proposed a greedy *k*-means algorithm to select a set of canonical views to form a scene summary, only based on visual features without exploiting the associated tags. Raguram and Lazebnik [9] selected iconic images to summarize general visual categories by using a joint GIST/pLSA clustering technique from both appearance and semantic aspects. They took the intersection of two independent clusters from the visual feature and the textual feature to get the final clustering, but the joint process was sequential instead of simultaneous. Fan et al. [10] first generated topic network for summarization, and then used a mixture-of-kernels and a representativeness-based image sampling algorithm to achieve image summarization. Li et al. [11] developed a multimedia application system, called as "Word2Image", by using the web image collections to translate a given word visually.

In some others research [11–14], the pairwise graph was adopted to describe the relationship between images and tags. They adopted the co-clustering technique to obtain image groups by using the texts surrounding the images. But they have only considered the association relations between images and tags.

In sum, most related approaches do not consider all the three types of relations over images and tags: image–image, tag–tag,

* Corresponding author.
*E-mail addresses:* minxianli.njust@gmail.com (M. Li), zhaochx@mail.njust.edu.cn (C. Zhao), jinhuitang@mail.njust.edu.cn (J. Tang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.02.050

and image–tag, simultaneously in the clustering procedure, so that visual exemplars are not depicted by semantic concepts well.

We use hypergraph to represent the complex and higher-order relations among images and tags. A hypergraph [15] is a generalization of a simple graph in which a set of vertices is defined as a weighted hyperedge [18], which is widely used in visual classification [19]. This characteristic enables hypergraphs to represent complex and higher-order relations which are difficult to be represented in traditional simple graphs. Spectral clustering was generalized from simple graphs to hypergraphs, while hypergraph embedding and transductive classification were further developed by spectral hypergraph clustering in [16]. In fact, spectral clustering [17] was usually utilized to solve the simple graph-based clustering problem [13], and its advantage over previous methods was verified in [18].

In this work we propose a novel hybrid image summarization approach by partitioning a hypergraph, which exploits the correlations among the low-level visual features, the correlation among the semantic tags, and the correlation between the visual features and the semantic tags, simultaneously. Moreover, the developed hypergraph partition method generates the exemplars that can effectively represent both the visual and semantic contents of the obtained clusters. Compared with prior work, the proposed approach is capable of exploiting both the correlations within images and tags, respectively, as well as the correlations between images and tags, to make the summarization results more reasonable.

The remainder of this paper is organized as follows. In Section 2, we detail how we create the hypergraph model according to the correlations within images and tags and the correlations between images and tags, and then present the hypergraph spectral partition algorithm. Moreover, we describe the representativeness-based visual and semantic exemplars selection algorithm. In Section 3 experimental details and results are reported. Finally, Section 4 concludes this paper.

## 2. The proposed approach

Given a query, image search engines or photo-sharing web sites usually generate an unstructured images collection and a large variety of text information associated with the images. We first represent the homogeneous correlations within images and tags, together with the heterogeneous correlations between images and tags, as a hypergraph. The hyperedge weights of these correlations are calculated according to the links between images and tags, and the similarities of inter-images and inter-tags. We then adopt a spectral hypergraph partition method to partition the image collection into several groups to ensure that the images in each group are both visually and semantically consistent. The hypergraph partition method is an extension of the simple graph spectral partition method, by associating each hypergraph with a natural random walk and then using the normalized cut approach. In the summarization, we represent each cluster group with an image exemplar and its associated representative tags. The image exemplars and associated representative tags are all found by using two representativeness score functions, which are also defined based on the feature co-occurrence, tag co-occurrence, and image–tag co-occurrence.

The proposed approach is superior over prior work because (1) the hypergraph framework takes advantage of homogeneous correlations with images and tags, respectively, as well as heterogeneous relations between them, this characteristic enhances the summarization performance; and (2) we take both visual and semantic representativeness into count to select exemplars, so that the image exemplars and tag exemplars are more representative for each cluster.

### 2.1. Hypergraph based clustering

#### 2.1.1. Preliminaries on hypergraph

Let $V$ represent a finite set of vertices and $E$ represent a family of subsets of $V$ such that $U_{e\in E}=V$. Then a *weighted hypergraph* can be denoted as $G(V,E,w)$, with the vertex set $V$ and the hyperedge set $E$, and each *hyperedge* $e$ is assigned a positive weight $w(e)$. The *degree of vertex* $v\in V$ is defined as $d(v)=\sum_{e\in E|v\in e}w(e)$ and the degree of hyperedge $e\in E$ is defined to be $\delta(e)=|e|$, where $|e|$ denotes the cardinality of $e$. A hypergraph $G$ can be represented by a $|V|\times|E|$ matrix $H$ called *incidence matrix* with entries $h(v,e)=1$ if $v_H\in e_H$ and zero otherwise. According to the definition of $H$, $d(v)=\sum_{e\in E}w(e)h(v,e)$ and $\delta(e)=\sum_{v\in V}h(v,e)$. Let us use $D_v$ and $D_e$ denote the diagonal matrices containing the vertex and hyperedge degrees, respectively, and let $W$ denote the diagonal matrix containing the weights of hyperedges. Then the *adjacency matrix* $A$ of hypergraph $G$ is defined to be $A=HWH^T-D_v$, where $H^T$ is the transpose of $H$ [16].

According to the above definition, different types of hyperedges may contain different types of vertices. The hypergraph with multiple types of vertices and hyperedges is called *unified hypergraph* [20]. Suppose a unified hypergraph has $m$ types of vertices and $n$ types of hyperedges, the vertex set of the $i$th type is denoted by $V_i$ and the hyperedge set of the $j$th type is denoted by $E_j$. So the vertex set and hyperedge set are defined by $V=U_{i=1}^{m}V_i$ and $E=U_{j=1}^{n}E_j$. Using the same notation in the unified hypergraph, we can describe the relations between images and their corresponding visual and semantic features.

#### 2.1.2. Construction of hypergraph

Given an image set $I=\{I_1,I_2,\ldots,I_n\}$, a corresponding tag set $T=\{T_1,T_2,\ldots,T_n\}$, where $n$ is the size of the image collection. Different kinds of features are extracted from images and the corresponding tag lists:

1. Visual features: Many of low-level visual features can be used, such as color histogram, texture and edge distribution. We extract $k$ different kinds of visual features denoted by $\mathcal{F}=\{F_i\}_{i=1}^{k}=\{[f_1^{(i)},f_2^{(i)},\ldots,f_n^{(i)}]\}_{i=1}^{k}$, where $f_j^{(i)}$ denotes the $i$th visual feature of the $j$th image in $\mathcal{I}$. We use three types of low-level features extracted from these images, including 64-D color histogram, 73-D edge direction histogram, and 128-D wavelet texture, while we use angle cosine as the image similarity.
2. Semantic features: The semantic features refer to the tags associated with images. Given a tag set $\mathcal{T}$, we extract the tag information $T_i=\{W_1^{(i)},W_2^{(i)},\ldots,W_{m_i}^{(i)}\}$ of each image, $m_i$ denotes the number of tags of $T_i$.

Based on these, we will introduce our hypergraph representation to model the homogeneous and heterogeneous relations among social images through hyperedges.

In the proposed unified hypergraph representation, a vertex set $V^I$ denotes image in a unified hypergraph, while a vertex set $V^T$ denotes the textual tags, and homogeneous and heterogeneous hyperedges are introduced to represent the various similarities and relationships between these vertices.

1. Homogeneous hyperedges
   The homogeneous hyperedges are conducted to describe similarities between homogeneous vertices. In this paper, there are two kinds of homogeneous hyperedges, one is utilized

to connect image vertices of $V^I$, which is called $E^I$, and the other is utilized to connect tag vertices of $V^T$, which is called $E^T$.

For each image vertex, we add $E^I$ hyperedge to connect it with its $k$-nearest neighbor image vertices, using different visual features respectively. In our experiment, $k$ is set to 400. In addition, we set the weight of each $E^I$ hyperedge with the average distance of its neighbors, which assumes that the closer its neighbors are, the higher weight the hyperedge will have.

For each tag vertex, we add $E^T$ hyperedge to connect it with its $k$-nearest neighbor tag vertices by joint probability. We first compute the frequency of co-occurrence of two tags, and take it as the distance of two tag vertices. In our experiment, $k$ is set to 100. Additionally, we set the weight of each $E^T$ hyperedge with the average distance of its neighbors.

2. Heterogeneous hyperedges

The heterogeneous hyperedges are conducted to represent the complex relationships between images and tags induced by social media, which connect each tag vertex with image vertices annotated by this tag. The weights of these kind hyperedges are set to 1.

#### 2.1.3. Hypergraph partition

Hypergraph partition algorithms can be divided into two categories [21]. One category intends to obtain a simple graph constructed from the original hypergraph, follow by partitioning the vertices by spectral clustering techniques. These approaches include clique expansion and star expansion [22], Rodriquez's Laplacian [23], etc. The other category defines a hypergraph "Laplacian" by using the analogies from the simple graph Laplacian. Characteristic methods in this category include Zhou's normalized Laplacian [16], Bolla's Laplacian [24], etc. [25] has verified that the above-mentioned methods are very close to each other in fact and they are equivalent under specific conditions. Specially, the hypergraph partition algorithm proposed in [16] is efficient and simple for implementation. In this paper, we adopt this algorithm to partition the hypergraph, which is presented as follows:

Given a vertex subset $S \subset V$, $S^c$ is denoted as the complement of $S$. The hyperedge boundary $\partial S$ is a hyperedges set to partition the hypergraph $G$ into two parts, $S$ and $S^c$ [16], and it can be defined as $\partial S := \{e \in E | e \cap S \neq \emptyset, \quad e \cap S^c \neq \emptyset\}$. Then, a two-way hypergraph partition could be defined as

$$H\mathrm{cut}(S,S^c) := \sum_{e \in \partial S} w(e) \frac{|e \cap S||e \cap S^c|}{\delta(e)}, \tag{1}$$

where $\delta(e)$ is the degree of the hyperedge $e$ defined in Section 2.1.1. The definition of the *hypergraph partition* given above can be assumed as a weighted sum of all hyperedges weights in $\partial S$. The two-way normalized hypergraph partition can be defined to avoid the bias of unbalanced partitioning:

$$NH\mathrm{cut}(S,S^c) := H\mathrm{cut}(S,S^c)\left(\frac{1}{\mathrm{vol}(S)} + \frac{1}{\mathrm{vol}(S^c)}\right), \tag{2}$$

where $\mathrm{vol}(S)$ is the volume of $S$, i.e., $\mathrm{vol}(S) = \sum_{v \in S} d(v)$ [21]. The form of Eq. (2) is very similar to the normalized cut [26] .

The combinatorial optimization problem given by Eq. (2) is NP-complete problem and it can be relaxed into a real-valued optimization problem as follows [16]:

$$\arg\min_{p \in \mathbb{R}^{|V|}} \sum_{e \in E} \sum_{\{u,v\} \subseteq e} \frac{w(e)}{\delta(e)} \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2 = \arg\min_{f \in \mathbb{R}^{|V|}} 2f^T \Delta f, \tag{3}$$

where $f$ is a label vector and the matrices $\Delta = I - D_v^{-1/2} HWD_e^{-1} H^T D_v^{-1/2}$, where $I$ denotes the identity matrix.

$\Delta$ is positive semi-definite, which is called the hypergraph Laplacian matrix. The theoretical solution of above optimization problem is the eigenvector of $\Delta$ associated with its smallest nonzero eigenvalue [16].

### 2.2. Hybrid summarization

Through hypergraph partition algorithm, the images and tags are partitioned into several groups. Our goal is to give a few exemplars to represent the image set semantically and visually. In other words, we need to select the most representative images and tags in each group, respectively. In the proposed algorithm, we first select semantic exemplars in each cluster, because we can use the results to help us select visual exemplars more accurately at the semantic level.

#### 2.2.1. Semantic exemplar selection

After hypergraph partition, the tags have been partitioned into several different groups. We select tag exemplars in each cluster in term of two criteria: (1) *High frequency*: More times the tags appear in a cluster, they are more representative to describe the cluster semantically. (2) *High coverage*: Due to the co-occurrence relationship between the tags, the tags with high frequency may derive from the same images subset in a cluster. Low coverage makes the semantic exemplars not representative. For describing the cluster more representatively, the semantic exemplars need cover the images as widely as possible.

As above-mentioned rules, a representativeness score $s(t,c)$ that measures how well tag $t$ describes the cluster $c$ which it belongs to can be defined as:

$$s(t,c) = \frac{1}{N_c} \sum_{i=1}^{N_c} \sum_{j=1}^{N_i^T} \phi(t), \tag{4}$$

where $N_c$ denotes the amount of the image in cluster $c$, $N_i^T$ denotes the amount of the tags associated with image $I_i$ in cluster $c$, and $\phi(t)=1$ if $t \in T_i$ and zero otherwise, which $T_i$ denotes the tag list associated with image $I_i$ in cluster $c$.

Actually, $s(t,c)$ defines the frequency of tag $t$. However, the tags with higher frequency do not mean that they can describe the cluster better, because they may all derive from the same images subset due to the co-occurrence relationship. Thus, we can redefine $s(t,c)$ as:

$$s(t,c) = \begin{cases} \frac{1}{N_c} \sum_{i=1}^{N_c} \sum_{j=1}^{N_i^T} \phi(t) & \text{if} \quad C(t, \tilde{t}) < T \\ 0 & \text{else} \end{cases}, \tag{5}$$

$$\tilde{t} = \arg\ \max\{s(t',c),\ t' \in K(t)\}, \tag{6}$$

where $K(t)$ is denoted as the $k$ nearest co-occurrence tags of $t$, and then $\tilde{t}$ is defined as the maximum $s(t',c)$ in $K(t)$. $C(t, \tilde{t})$ is termed as the co-occurrence relation between tag $t$ and tag $\tilde{t}$, $T$ is the threshold. In our experiment, $k$ is set to 10, and $T$ is set to 0.5. Then the tag exemplars $T_E$ can be selected by ranking the representativeness score $s(t,c)$ of each tag.

#### 2.2.2. Visual exemplar selection

There are also two criteria for choosing the most representative images in our representativeness-based image sampling technique: (1) *Image clustering*: Through the proposed hypergraph-based image clustering algorithm, we can obtain a good global distribution structure (i.e., image clusters and their relationships) for large amounts of images under the same group, which are similar in low-level visual features. Thus we achieve the adaptive image sampling by selecting the most representative images to summarize the visually-similar images in the same cluster. (2) *Tag clustering*: The hypergraph model considers three types of relations over

images and tags: image–image, tag–tag, and image–tag simultaneously. That said, the images in the same group are similar not only visually but also semantically. Thus adaptive image sampling can be achieved by selecting the most representative images to summarize the semantically-similar images in the same cluster.

For the images in the same cluster, we can define a score to describe the representativeness of each image, which depends on their closeness with the cluster centers visually and semantically. The representativeness score $\rho(x,y)$ for the given image $I_i^c$ with the visual features $f_i$ and the semantic features $T_i$ can be defined as:

$$\rho(f_i,T_i,c) = \alpha e^{-\left(f_i-\mu_c^f\right)^T\left(f_i-\mu_c^f\right)} + (1-\alpha)e^{\varphi(T_i,T_E)} \quad (7)$$

where $\mu_c^f$ is the center of visual features of the cluster c, $\varphi(T_i,T_E)$ denotes that how well the semantic feature $T_i$ matches with the semantic exemplars $T_E$, and $\alpha$ is the factor which is used to balance the weight of visual features and semantic features. Thus the images, which are closer to the cluster centers of visual features and have more semantic exemplars, have larger values of $\rho(\cdot,\cdot,\cdot)$. The images in the same cluster $c$ can be ranked according to their representativeness scores, and the most representative images with larger values of $\rho(\cdot,\cdot,\cdot)$ can be selected to generate the image exemplars based on the visual and semantic similarity.

## 3. Evaluation

To evaluate the performance of the proposed hybrid image summarization approach, we compare it with a baseline method.

### 3.1. Dataset

We use 269,648 images and the associated tags from NUS-WIDE database [27], which are crawled from the popular photo-sharing web site Flickr, with a total of 5018 unique tags. Some concepts are used for querying, including flowers, beach, building, dog, cat, plants, mountain, river, and so on. It is noted that these concepts are not abstract and familiar for people, so that the experimental evaluation can be accurate.

### 3.2. Summarization results comparison

The hybrid summarization results are presented on four representative subsets of images from NUS-WIDE database, "flowers", "beach", "wedding", and "mountain". We crawled top 4782, 8092, 1576, and 3493 images from each subset, respectively.

We compare the performance of the proposed approach with a baseline method—affinity propagation (AP) [28]. AP is an algorithm to spontaneously select a good subset of exemplars for a whole set of data points, by considering all data points as candidate exemplars such that they can represent the image collection very well. Because the AP algorithm can only propagate one type relationship, we select the visual relationship due to its importance in image collections. More specifically, we extract three low-level visual features, including 64-D color histogram, 73-D edge direction histogram, and 128-D wavelet texture. Then, we use angle cosine of the visual features as the image similarity, which is propagated in AP algorithm.

The visual and sematic summarization result for the query "flowers" is demonstrated in Fig. 1, and the summarization comparison results are depicted in Figs. 2–4. Our results look appealing in both visual and semantic performances, and the visual summary can capture the semantic meaning well. The AP algorithm cannot get competitive performance because it has no ability to exploit visual and semantic features simultaneously.

### 3.3. User study

We conduct two user studies to evaluate the proposed hybrid summarization approach. The first user study is to evaluate the effectiveness of the proposed approach, and the second user study is to compare our approach with affinity propagation algorithm. These two user studies are both carried on all four subsets.

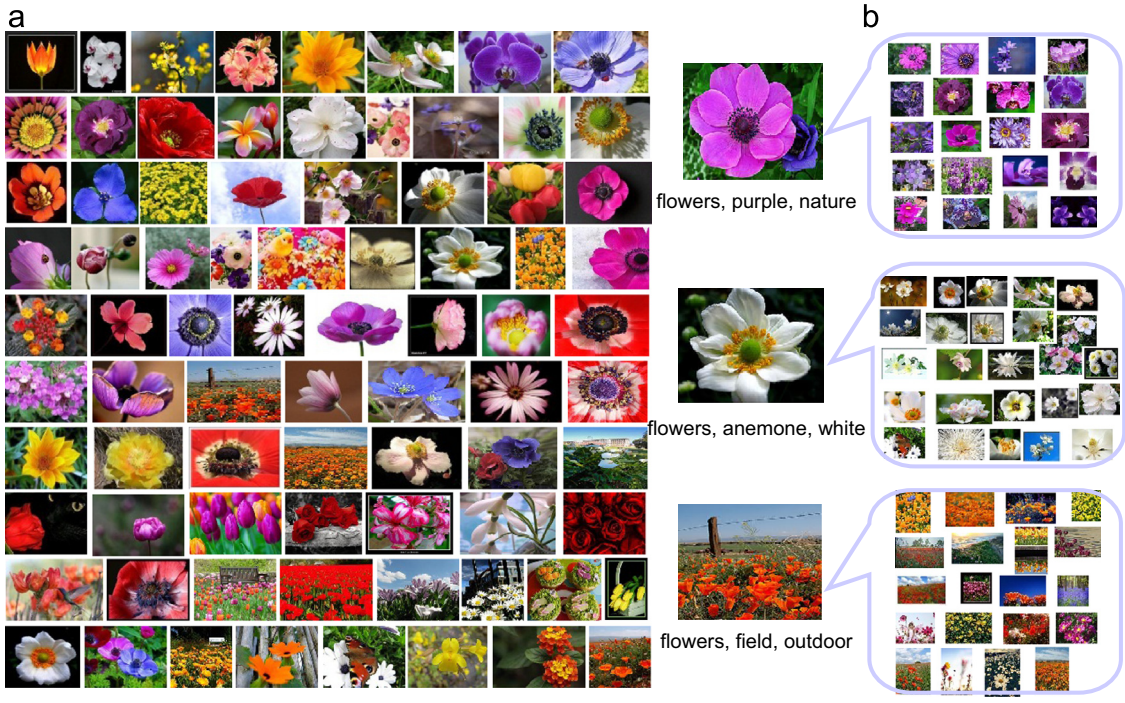


**Fig. 1.** An example of summary for a real-world image dataset crawled from Flickr by the proposed hybrid summarization scheme: (a) randomly selected images and their associated texts for "flowers", (b) visual and semantic summarization.

We have invited 20 people to join the user study, including 12 men and 8 women. These participants cover different background, including graduate students, researchers, educator, and business man. Their ages range from 23 to 42.

### 3.3.1. *Evaluation of effectiveness*

To evaluate the effectiveness of the proposed approach, participants were asked to answer the following four questions, by giving a rating score of between 1 (bad) and 10 (excellent):

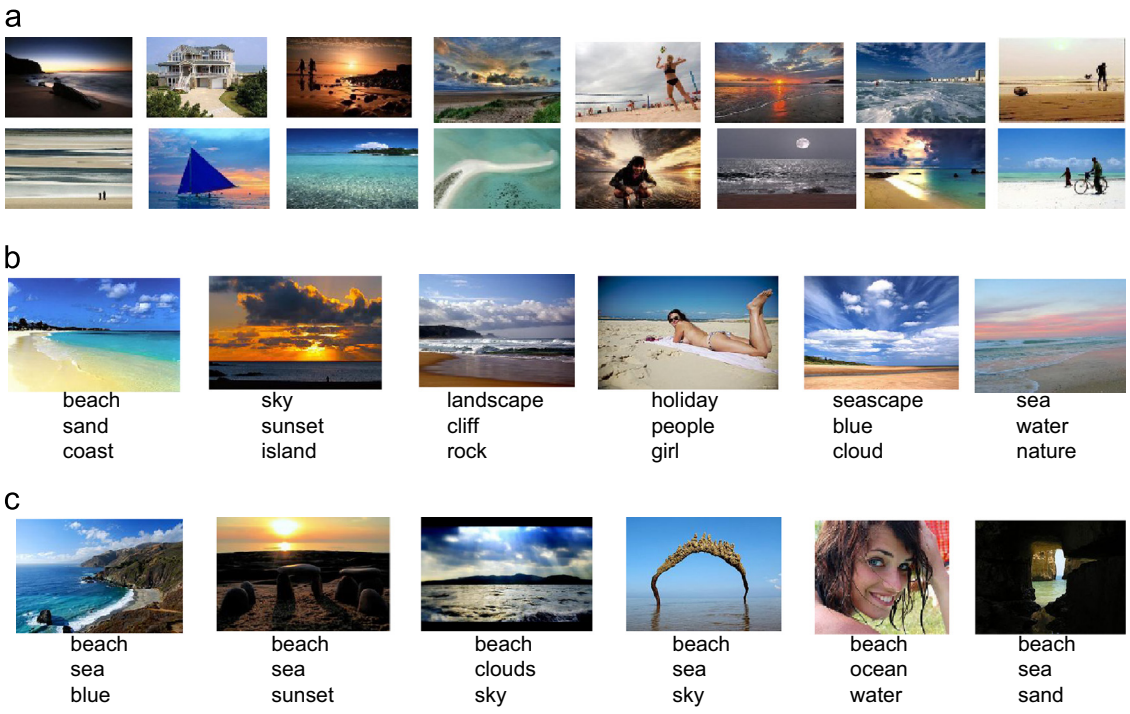


**Fig. 2.** The "beach" results: (a) the examples of original images, (b) results of our approach, (c) results of AP.


**Fig. 3.** The "wedding" results: (a) the examples of original images, (b) results of our approach, (c) results of AP.

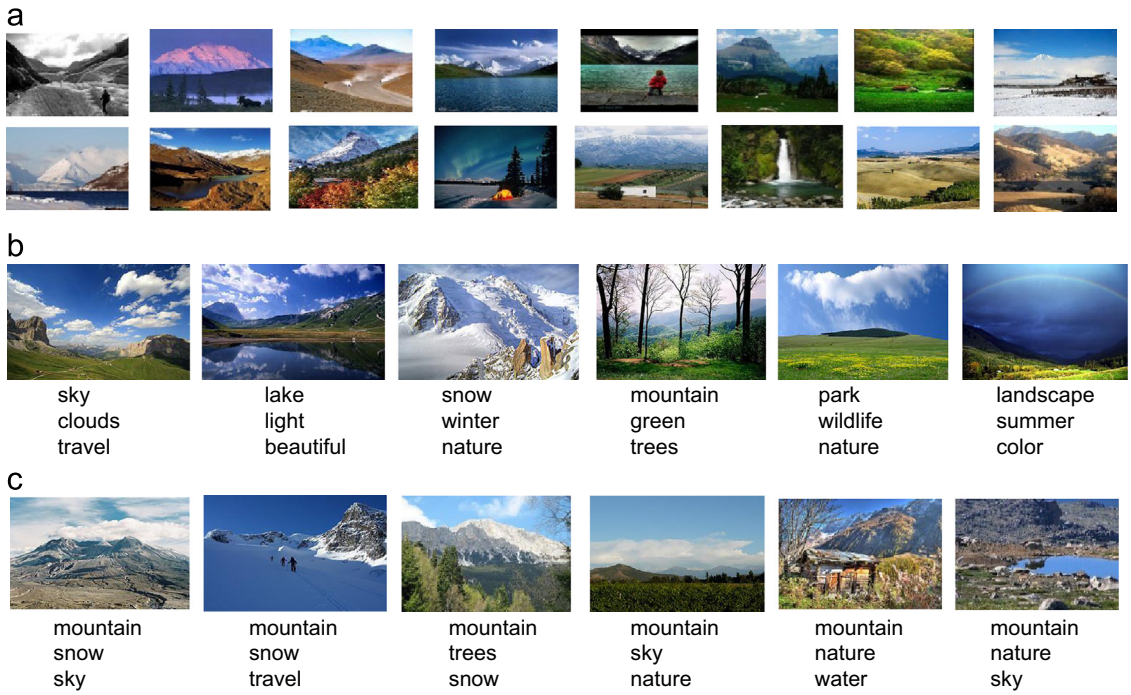


**Fig. 4.** The "mountain" results: (a) the examples of original images, (b) results of our approach, (c) results of AP.

1. Selected image exemplars should represent the corresponding image group. Based on the set of image exemplars given, how do you evaluate the performance of the proposed approach on representativeness of image exemplars?
2. Selected image exemplars should describe different aspects of the corresponding image group. Based on the set of image exemplars given, how do you evaluate the performance of the proposed approach on diversity of image exemplars?
3. Selected tag exemplars should represent the corresponding image group. Based on the set of tag exemplars given, how do you evaluate the performance of the proposed approach on representativeness of tag exemplars?
4. Selected tag exemplars should describe different aspects of the corresponding image group. Based on the set of tag exemplars given, how do you evaluate the performance of the proposed approach on diversity of tag exemplars?

For the first question, participants were required to give a score from 1 to 10. Fig. 5 describes the evaluation result, which shows that 85% of users think that the selected images are representative for the corresponding image group (with scores of $\geq 6$).

Fig. 6 shows the evaluation result from all 20 participants for the second question. A higher score means more diversity between image exemplars in the final results. The result shows that 90% of users think that the selected images are diverse for the corresponding image group (with scores of $\geq 6$).

For the third question, Fig. 7 describes the evaluation result, which shows that 80% of users think that the selected tags are representative for the corresponding image group (with scores of $\geq 6$).

For the fourth question, a higher score means more diversity between tag exemplars in the final results. Fig. 8 shows that 85% of users think that the selected tags are diverse for the corresponding image group (with scores of $\geq 6$).

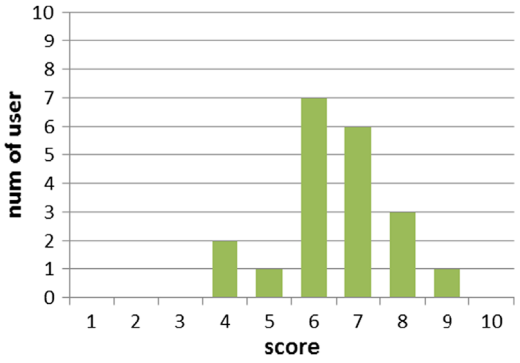


**Fig. 5.** Evaluation on representativeness of image exemplars.

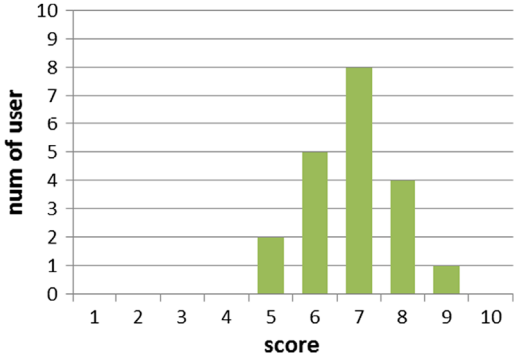


**Fig. 6.** Evaluation on diversity of image exemplars.

*3.3.2. Comparison with a baseline method*

In this overall evaluation, comparisons between the effectiveness of the proposed approach and affinity propagation were conducted. Participants were asked to provide a score of between $-5$ and 5 ($-5$: the worst; $-4$: much more worse; $-3$: much worse; $-2$: worse; $-1$: a little worse; 0: similar; 1: a little better;

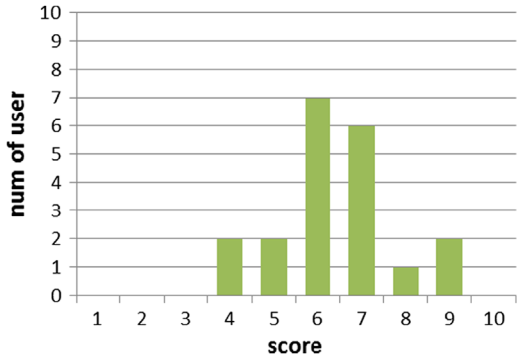


**Fig. 7.** Evaluation on representativeness of tag exemplars.

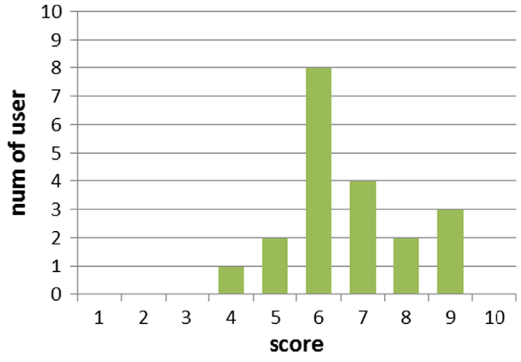


**Fig. 8.** Evaluation on diversity of tag exemplars.

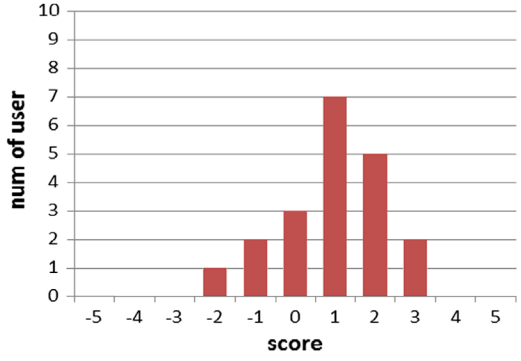


**Fig. 9.** Visual comparison results.

2: better; 3: much better; 4: much more better; 5: the best) for two questions:

1. How is the representativeness and diversity of the image exemplars returned by the proposed approach compared with affinity propagation, respectively?
2. How is the representativeness and diversity of the tag exemplars returned by the proposed approach compared with affinity propagation, respectively?

The evaluation results are shown in Figs. 9 and 10.

Fig. 9 shows the visual comparison between the proposed approach and AP. As shown in the results, all participants agreed that our image exemplars are better than AP visually, while most of the participants (70%) voted the proposed approach to be much better than AP (with scores of $\geq 1$).

Fig. 10 shows the semantic comparison between the proposed approach and AP. As indicated from the data, all participants

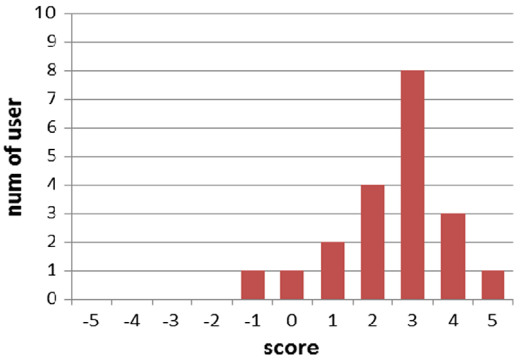


**Fig. 10.** Semantic comparison results.

agreed that our tag exemplars are better than AP semantically, while most of the participants (90%) voted the proposed approach to be much better than AP (with scores of $\geq 1$). The reason is that AP algorithm is unable to get tag exemplars simultaneously, while it partitions the image set to get image exemplars.

## 4. Conclusion

In this paper, we present a novel hybrid image summarization scheme to manage image collections, which bases on a hypergraph model. We first extract the low-level visual features and the textual tags from an internet image collection, and then construct hyperedges using three useful relations, including two homogeneous relations within images and tags and a heterogeneous association relation between images and tags. The hyperedges are defined as a set formed by each vertex and its $k$-nearest neighbors, and their weights are calculated by the sums of corresponding pairwise affinities. We formulate the image clustering as the problem of hypergraph partition and use a generalized spectral clustering technique to solve it. After clustering, we design two representativeness score functions to select the visual exemplars and semantic exemplars, respectively. The experimental results of the proposed method are effective on a real internet image collection. In the future, the effective user feedback technique may be integrated into our system to boost the performance [29].

## Acknowledgments

This work was partially supported by the Natural Science Foundation of China (NSFC) under Grant 61103059, the National Basic Research Program of China (973 Program) under Grant 2012CB316304, the NSFC under Grant 61173104, the Natural Science Foundation of Jiangsu Province for Distinguished Young Scholars under Grant BK2012033, The Natural Science Foundation of Jiangsu Province under Grant BK2011700, Research Fund for the Doctoral Program of Higher Education of China (RFDP) under Grant 20113219120022.

## References

[1] 〈http://www.flickr.com/〉(2011-10-01).
[2] 〈http://picasa.google.com/〉(2011-10-01).
[3] J. Tang, R. Hong, S. Yan, T.S. Chua, G.J. Qi, R. Jain,Image, Annotation by knn-sparse graph-based label propagation over noisily-tagged web images, ACM Trans. Intel. Syst. Technol. 2 (2) (2011).
[4] J. Tang, H. Li, G.J. Qi, T.S. Chua, Image annotation by graph-based inference with integrated multiple/single instance representations, IEEE Trans. Multimedia 12 (2) (2010).

[5] R. Clough, H. Joho, M. Sanderson, Automatically organizing images using concept hierarchies, in: SIGIR Workshop on Multimedia Information Retrieval, 2005.
[6] P. Schmitz, Inducing Ontology from Flickr tags, in: Proceedings of the Workshop on Collaborative Tagging at World Wide Web, 2006.
[7] A. Jaffe, M. Naaman, T. Tassa, M. Davis, Generating Summaries for Large Collections of Geo-Referenced Photographs. in:Proc. World Wide Web, 2006, pp. 853–854.
[8] I. Simon, N. Snavely, S.M. Seitz, Scene Summarization for online image collections, in: Proceedings of the IEEE 11th International Conference on Computer Vision, 2007, pp. 1–8.
[9] R. Raguram S. Lazebnik, Computing iconic summaries for general visual concepts, in: Workshop on Internet Vision at Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[10] J. Fan, Y. Gao, H. Luo, D.A. Keim, Z. Li, A novel approach to enable semantic and visual image summarization for exploratory image search, in: Proceedings of the Multimedia Information Retrieval, 2008, pp. 358–365.
[11] H. Li, J. Tang, G. Li, T.S. Chua, Word2Image: towards visual interpreting of words, in: Proceedings of the 16th ACM International Conference on Multimedia, 2008.
[12] M. Rege, M. Dong, F. Fotouhi, Co-clustering image features and semantic concepts, in: Proceedings of the IEEE International Conference on Image Processing, 2006, pp. 137–140.
[13] B. Gao, T.Y. Liu, T. Qin, X. Zheng, Q. Cheng, W.Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts, in: Proceedings of the ACM International Conference Multimedia, 2005, pp. 112–121.
[14] M. Rege, M. Dong, J. Hua, Graph Theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering, in: Proceedings of the 17th International Conference on World Wide Web for Efficient Web Image Clustering, 2008, pp. 317–326.
[15] C. Berge, Hypergraphs, North-Holland, Amsterdam, 1989.
[16] D. Zhou, J. Huang, B. SchÄolkopf, Learning With hypergraphs: clustering, classification, and embedding, in: Proceedings of the Conference on Advances in Neural Information Processing Systems, 2007, pp. 1601–1608.
[17] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, in: Proceedings of the Advances in Neural Information Processing Systems, 2001.
[18] T. Tuytelaars, C.H. Lampert, M.B. Blaschko, W. Buntine, Unsupervised, Object discovery: a comparison, Int. J. Comput. Vision (2009).
[19] J. Tang, X.S. Hua, et al., Correlative linear neighborhood propagation for video annotation, IEEE Trans. Syst. Man Cybern. Part B Cybern. 39 (2) (2009).
[20] J. Bu, S. Tan, Ch. Chen, et al., Music recommendation by unified hypergraph: combining social media information and music content, in: Proceedings of the International Conference on Multimedia, 2010, pp. 391–400.
[21] Y. Huang, Q. Liu, F. Lv, Y. Gong, D.N. Metaxas, Unsupervised image categorization by hypergraph partition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (6) (2011) 1266–1273.
[22] J.Y. Zien, M.D.F. Schlag, P.K. Chan. Multi-level spectral hypergraph partitioning with arbitrary vertex sizes, in: Proceedings of the IEEE International Conference Computer Aided Design, 1996, pp. 201–204.
[23] J. Rodrequez, On the Laplacian spectrum and walk-regular hypergraphs, Linear Multilinear Algebra 51 (2003) 285–297.
[24] M. Bolla, Spectra, Euclidean Representations and Clustering of Hypergraphs, Discrete Math., 1993, Proc.
[25] S. Agarwal, K. Branson, S. Belongie, Higher order learning with graphs, in: Proceedings of the International Conference Machine Learning, 2006.
[26] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
[27] T.S. Chua, J. Tang, et al. NUS-WIDE: a real-world web image database from National University of Singapore, in: ACM International Conference on Image and Video Retrieval, 2009.
[28] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.
[29] J. Tang, Z.J. Zha, D. Tao, T.S. Chua, Semantic-gap oriented active learning for multi-label image annotation, IEEE Trans. Image Process. 21 (4) (2012) 2354–2360.

**Minxian Li** was born in Jiangsu Province, China in 1983. He is now a Ph.D. student in the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing. His research focuses on social multimedia analysis and retrieval.

**Chunxia Zhao** received the B.E., M.S. and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 1985, 1988, 1998, respectively, both in the Department of Electrical Engineering and Computer. She is a professor in the Department of Computer Science, Nanjing University of Science and Technology. She had a stay of one year as research assistant at the Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong from 1997 to 1998. Her current interests are in the areas of robots, computer vision, and pattern recognition.

**Jinhui Tang** received the B.E. and Ph.D. degrees from University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively, both in the Department of Electronic Engineering and Information Science. Since July 2008, he has been a Research Fellow in the School of Computing, National University of Singapore. Now he is a professor in the Department of Computer Science, Nanjing University of Science and Technology. His current research interests include content-based image retrieval, video content analysis, and pattern recognition. Prof. Tang is a member of the Association for Computing Machinery. He is a recipient of the 2008 President Scholarship of Chinese Academy of Science, and a co-recipient of the Best Paper Award in ACM Multimedia 2007.