Hybrid Affinity Propagation[*]

Jingdong Wang[†] Hao Xu[‡] Xian-Sheng Hua[†] Shipeng Li[†]
[†]Microsoft Research Asia [‡]University of Science and Technology of China
{jingdw, xshua, spli}@microsoft.com xuhao657@ustc.edu

Abstract

In this paper, we address a problem of managing tagged images with hybrid summarization. We formulate this problem as finding a few image exemplars to represent the image set semantically and visually, and solve it in a hybrid way by exploiting both visual and textual information associated with images. We propose a novel approach, called homogeneous and heterogeneous message propagation ($\mathrm{H^2MP}$). Similar to the affinity propagation (AP) approach, $\mathrm{H^2MP}$ reduce the conventional *vector* message propagation to *scalar* message propagation to make the algorithm more efficient. Beyond AP that can only handle homogeneous data, $\mathrm{H^2MP}$ generalizes it to exploit extra heterogeneous relations and the generalization is non-trivial as the reduction to scalar messages from vector messages is more challenging. The main advantages of our approach lie in 1) that $\mathrm{H^2MP}$ exploits visual similarity and in addition the useful information from the associated tags, including the associations relation between images and tags and the relations within tags, and 2) that the summary is both visually and semantically satisfactory. In addition, our approach can also present a textual summary to a tagged image collection, which can be used to automatically generate a textual description. The experimental results demonstrate the effectiveness and efficiency of the proposed approach.

1 Introduction

The increasing development of image search engines, photo-sharing web sites, and desktop photo management tools, has made people easily access a large amount of images. However, image collections are usually unorganized, which makes finding desired photos and quick overview of an image collection very difficult. This unstructured nature of image collections has attracted great effort on computing visual summaries. On the other hand, most image collections are provided with rich text information, and such image collections are called tagged image collections in this paper. For example, images on Flickr are titled, tagged, and commented by users. Images from the Web are often associated with surrounding texts. The text information usually reflects the semantic content of images and is helpful for summarization.

In this paper, we address the image management task through a hybrid summarization scheme. The key is to find the summary in a hybrid way to exploit both visual and textual information. An example is shown in Fig. 1. Given rich tag information associated with images, there are three useful relations from images and tags: two homogeneous relations within images and tags, including image similarity and tag similarity,

[*] A short version appeared in [19]

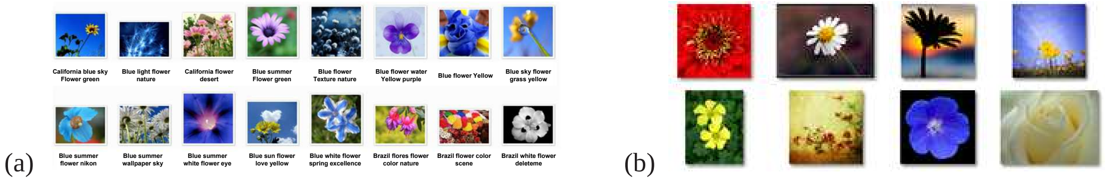

(a)
California blue sky Flower green
Blue light flower nature
California flower desert
Blue summer Flower green
Blue flower Texture nature
Blue flower water Yellow purple
Blue flower Yellow
Blue sky flower grass yellow
Blue summer flower nikon
Blue summer wallpaper sky
Blue summer white flower eye
Blue sun flower love yellow
Blue white flower spring excellence
Brazil flores flower color nature
Brazil flower color scene
Brazil white flower deleteme
(b)

Figure 1: An example of a visual summary for an image collection. (a) shows randomly selected images and their associated texts from the input, a set of tagged images, and (b) shows its summary identified by our hybrid summarization scheme.

and one heterogeneous relation between images and tags, e.g., their association relations. We propose a hybrid summarization approach to find image exemplars through investigating all three relations together including the information from the associated tags, i.e., association relations between images and tags and relations within tags so that the summary is both visually and semantically satisfactory.

1.1 Related work

Most existing image sharing web sites present an overview of an image collection by showing the top images (e.g., Flickr group [4]), which obviously does not present a good summary, or allowing consumers to manually select images (e.g., Picassa web album [1]), which is inconvenient for consumers particularly in a large number of images.

Rother et al. [15] summarize a set of images with a "digital tapestry". They synthesize a large output image from a set of input images, stitching together salient and spatially compatible blocks from the input images. Wang et al. [18] create a "picture collage", a 2D spatial arrangement of the images in the input set chosen to maximize visible salient regions. These works do not address the problem of selecting the set of images to appear in the summary.

Recently, there are a few works to deal with the selection problem. Simon et al. [17] selects a set of images using the greedy k-means algorithm, by examining the distribution of images to select a set of canonical views only based on visual features without exploiting the associated tags. Raguram and Lazebnik [13] select iconic images to summarize general visual categories using a simple joint clustering technique from both appearance and semantic aspects. It first obtained two independent clusters from the visual feature and the textual feature, respectively, and then takes their intersection to get the final clustering, but the joint process is obtained sequentially instead of simultaneously. Surrounding texts are limitedly exploited for image grouping [9, 14] by considering the association relations between words and images using the co-clustering technique, but without investigating interior relations over tags.

Image summarization is also studied in the information retrieval community. Clough et al. [5] construct a hierarchy of images using only textual caption data, and the concept of subsumption. Schmitz [16] uses a similar approach but relies on Flickr tags. Jaffe et al. [10] summarize a set of images using only tags and geotags. By detecting correlations between tags and geotags, they are able to produce tag maps, where tags and related images are overlaid on a geographic map at a scale corresponding to the range over which the tag commonly appears. All these approaches could be used to further organize the images. However, none of them exploits the visual information.

1.2 Our approach

In this paper, we present a hybrid summarization approach to find a few image exemplars to represent the image collection, which is both visually and semantically satisfactory. Toward this end, we propose an effective scalar hybrid message propagation scheme over images and tags, homogeneous and heterogeneous message propagation (H^{2}MP), to exploit simultaneously homogeneous relations within images and tags and heterogeneous relations between images and tags. It is beyond the affinity propagation algorithm [8] that only can handle homogeneous data, and H^{2}MP can effectively exploit the heterogeneous relations between images and tags as well as the interior relations within tags. Moreover, the reduction from vector messages to scalar messages is more challenging than AP because H^{2}MP contains additional heterogeneous relations. Besides, our approach is superior over existing co-clustering algorithms [6, 7] that only utilize the heterogeneous relations because 1) it directly obtains the exemplars instead of performing the necessary postprocess to find the centers followed by a clustering procedure and 2) it takes advantage of homogeneous relations with images and tags as well as heterogeneous relations between them.

1.3 Notation

Given a set of n images, $\mathcal{I} = \{I_1, I_2, \cdots, I_n\}$, a set of corresponding texts, $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_n\}$, $\mathcal{T}_k = \{W_1^k, W_2^k, \cdots, W_{m_k}^k\}$, and the union set of tags $\mathcal{W} = \{W_1, \cdots, W_m\} = \mathcal{T}_1 \cup \mathcal{T}_2 \cup \cdots \cup \mathcal{T}_n$, we aim to find a summary, a set of image exemplars, $\bar{\mathcal{I}} \in \mathcal{I}$. There are three types of relations over images and tags as depicted in Fig. 2. The heterogeneous relations between all pairs of associated images and tags are represented by the edges, $\mathcal{E}^R$. The homogeneous relations within images are represented by the edges, $\mathcal{E}^I$, and the similarity between a pair of images i and k is denoted by $s'_I(i, k)$. The homogeneous relations within tags are represented by the edges, $\mathcal{E}^W$, and the similarity between a pair of tags j and k is represented by $s'_W(j, k)$.

Suppose a set of image exemplars to be identified be denoted as $\bar{\mathcal{I}} = \{I_{c_1}, I_{c_2}, \cdots, I_{c_n}\}$, where $c_k \in \{1, 2, \cdots, n\}$ is the exemplar image index of image I_k, and $\mathbf{c} = [c_1 \ c_2 \ \cdots \ c_n]^T$ is a label vector over images. If such a label vector satisfies a valid constraint that an image should also serve as the exemplar of itself if it is an exemplar of any other image, it would uniquely correspond to a set of image exemplars. In other words, identifying the exemplars can be viewed as searching over valid labels.

2 Affinity propagation

To make our approach easily understood, we first review affinity propagation (AP) [8]. AP is an approach to find a good subset of exemplars for a whole set of homogeneous data points, by considering all data points as candidate exemplars such that they can represent the image collection very well, and mathematically it is formulated to maximize an objective function,

$$\mathcal{S}(\mathbf{c}) = E^I(\mathbf{c}) + V^I(\mathbf{c}), \tag{1}$$

where $E^I(\mathbf{c})$ is a fitting function to evaluate how well the image exemplars represent the other images, and written as

$$E^I(\mathbf{c}) = \sum\nolimits_{i=1}^{n} s'_I(i, c_i), \tag{2}$$

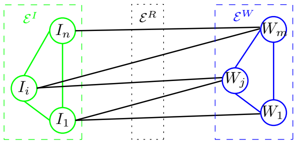


Figure 2: Heterogeneous graph over images and tags.

and $V^I(\mathbf{c})$ is a valid configuration function to constrain that an image must select itself as its exemplar if it is selected as an exemplar of other data point, and it is formulated as

$$V^I(\mathbf{c}) = \sum\nolimits_{k=1}^{n} \delta_k(\mathbf{c}), \tag{3}$$

where

$$\delta_k(\mathbf{c}) = \begin{cases} -\infty, & \text{if } c_k \neq k \text{ but } \exists i : c_i = k \\ 0, & \text{otherwise}. \end{cases}$$

This objective function can be depicted by a factor graph over the variables $\{c_i\}$, and the function terms $\{\delta_i\}$ and $\{s_i\}$, which correspond to the subgraph in the dashed green box in Fig. 3. AP to maximize Eqn. (1) is a scalar message propagation algorithm and derived from the max-sum algorithm, which transits two vector-messages between c_i and δ_k. The message, $\rho_{i\rightarrow k}$, sent from c_i to δ_k, consists of n real numbers, with one for each possible value of c_i. The message, $\alpha_{i\leftarrow k}$, sent from δ_k to c_i, also consists of n real numbers. For simplicity, the subscript I may be dropped in the following presentation. The two messages are depicted in Figs. 4(a) and 4(b) neglecting the massages corresponding to the red edges, and formulated as follows,

$$\rho_{i\rightarrow k}(c_i) = s(i, c_i) + \sum\nolimits_{k':k'\neq k} \alpha_{i\leftarrow k'}(c_i), \tag{4}$$

$$\begin{aligned} \alpha_{i\leftarrow k}(c_i) = & \max_{h_1,\cdots,h_{i-1},h_{i+1},\cdots,h_n} [\sum\nolimits_{i':i'\neq i} \rho_{i'\rightarrow k}(h_{i'}) \\ & + \delta_k(h_1, \cdots, h_{i-1}, c_i, h_{i+1}, \cdots, h_n)]. \end{aligned} \tag{5}$$

As shown in [8], this vector-message propagation scheme can be reduced to a scalar-message propagation scheme between data points. There are two kinds of messages exchanged within image points. The "responsibility" $r(i, k)$, sent from data point i to data point k, which reflects how well k serves as the exemplar of i considering other potential exemplars for i, and the "availability" $a(i, k)$, sent from data point k to data point i, which reflects how appropriately i chooses k as its exemplar considering other potential points that may choose k as their exemplar. The messages are updated in an iterative way as

$$r(i, k) = s(i, k) - \max\nolimits_{i':i'\neq k}[s(i, i') + a(i, i')]. \tag{6}$$

$$a(i, k) = \begin{cases} \sum_{i':i'\neq k} \max(0, r(i', k)) & k = i, \\ \min[0, r(k, k) + \sum_{i':i'\neq i,k} \max(0, r(i', k))] & k \neq i. \end{cases} \tag{7}$$

3 Hybrid summarization

Affinity propagation has been shown to be very effective to find exemplars [8], but it can only handle homogeneous data points. In the following, we present a new hybrid message propagation approach, to generalize AP to heterogeneous data points, homogeneous and heterogeneous message propagation (H^{2}MP). H^{2}MP is applied to find image exemplars for a tagged image collection, with the advantages of exploiting not only

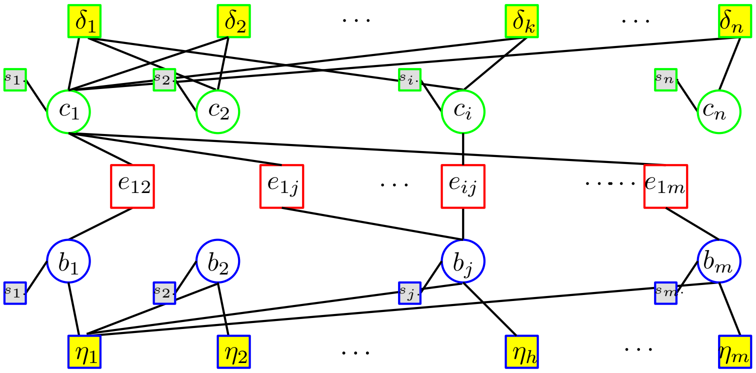


Figure 3: Factor graph for the overall objective function Eqn. (23). □ represents a function node, and ◯ represents a variable node.

the visual information of images but also heterogeneous relations between images and their associated tags and interior similarities within words in a hybrid way.

We exploit the associated tag information for image identification by augmenting image exemplar identification with tag exemplar identification and bridging image exemplars and tag exemplars according to association relations between images and tags. Thus, we define a label vector over tags $\mathbf{b} = [b_1\ b_2\ \cdots\ b_m]^T$ to represent the tag exemplars, i.e., W_j selects W_{b_j} as its exemplar. Basically, the proposed approach is modeled according to the following two properties: 1) these image and tag exemplars are good representatives of images and tags, respectively, and 2) these image and tag exemplars reflect their association relations. The second property investigates the association relations between images and tags and also serves as a bridge to make use of the relations within tags.

The first property concerns how well image and word exemplars represent the other images and tags if only the homogeneous relations are taken into consideration. For images, this is formulated as $E^I(\mathbf{c})$ in Eqn. (1), and for tags we can get a similar formulation,

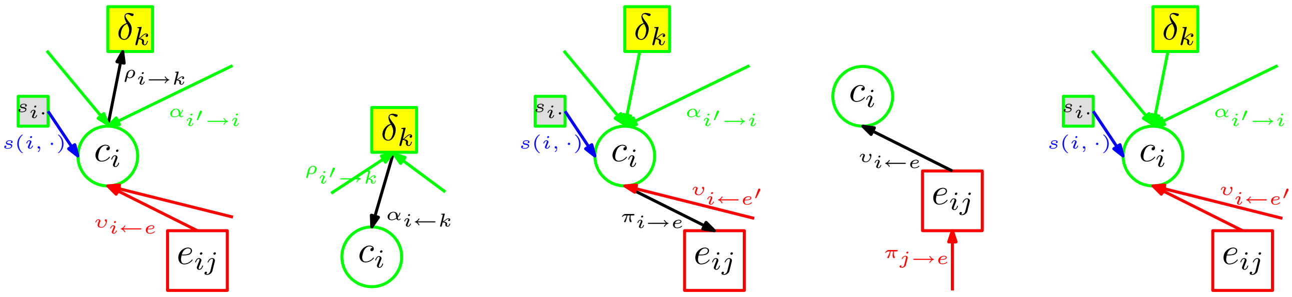
$$E^W(\mathbf{b}) = \sum\nolimits_{j=1}^{m} s'_W(j, b_j). \tag{8}$$

The second property essentially investigates the effect of the heterogeneous relations between images and tags on exemplar identification, and serves as a bridge to get help for image exemplar identification from tag information. We would like to assign different preferences for a pair of connected image i and tag j according to whether image i or tag j selects itself as its exemplar. This affect is formulated as a function $e_{ij}(c_i, b_j)$ over a pair of image and tag (i, j) and their exemplars (c_i, b_j), $e_{ij}(c_i, b_j)$. The whole affect function is written as follows,

$$R(\mathbf{c}, \mathbf{b}) = \sum\nolimits_{(i,j)\in\mathcal{E}^R} e_{ij}(c_i, b_j), \tag{9}$$

where $e_{ij}(c_i, b_j)$ aims to set different weights according to whether c_i is equal to i and whether b_j is equal to j,

$$e_{ij}(c_i, b_j) = \begin{cases} q(i,j), & c_i \neq i, b_j \neq j \\ \bar{q}(i,j), & c_i = i, b_j = j \\ p(i,j), & c_i = i, b_j \neq j \\ p(j,i), & c_i \neq i, b_j = j. \end{cases}$$



(a) $\rho_{i\rightarrow k}$ (b) $\alpha_{i\leftarrow k}$ (c) $\pi_{i\rightarrow e}$ (d) $\upsilon_{i\leftarrow e}$ (e) belief

Figure 4: Vector-valued messages.

In this paper, we instantiate this affect function similar to the Ising model based on the following aspects: if an image is selected as an exemplar, the tags linking this image should have larger probability to be selected as exemplars, and vice versa. In other words, we would assign a higher penalty for a pair of related image i and tag j when they do not select themselves as their exemplars simultaneously or do not select others as their exemplars simultaneously. Specifically, we set $p(i,j)$ and $p(j,i)$ to negative values and set $q(i,j)$ and $\bar{q}(i,j)$ to zero.

For $\mathbf{b}$, a similar valid constraint is defined as

$$V^W(\mathbf{c}) = \sum\nolimits_{k=1}^{m} \eta_k(\mathbf{b}), \tag{10}$$

and $\eta_k(\cdot)$ is defined similarly to $\delta_k(\cdot)$.

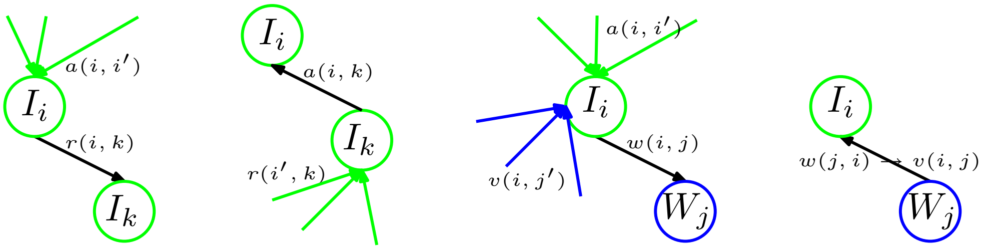
In summary, the overall objective function is as follows

$$\begin{aligned}\mathcal{S}(\mathbf{c},\mathbf{b}) =& \gamma_I(E^I(\mathbf{c}) + V^I(\mathbf{c})) + \gamma_W(E^W(\mathbf{b}) + V^W(\mathbf{b})) + R(\mathbf{c},\mathbf{b}) \\ =& \sum\nolimits_{i=1}^{n} \gamma_I s_I'(i,c_i) + \sum\nolimits_{k=1}^{n} \delta_k(\mathbf{c}) + \sum\nolimits_{j=1}^{m} \gamma_W s_W'(j,b_j) \\ & + \sum\nolimits_{k=1}^{m} \eta_k(\mathbf{b}) + \sum\nolimits_{(i,j)\in\mathcal{E}^R} e_{ij}(c_i,b_j) \\ =& \sum\nolimits_{i=1}^{n} s_I(i,c_i) + \sum\nolimits_{k=1}^{n} \delta_k(\mathbf{c}) + \sum\nolimits_{j=1}^{m} s_W(j,b_j) \\ & + \sum\nolimits_{k=1}^{m} \eta_k(\mathbf{b}) + \sum\nolimits_{(i,j)\in\mathcal{E}^R} e_{ij}(c_i,b_j),\end{aligned} \tag{11}$$

where γ_I and γ_W are balance weights, $s_I(i,c_i) = \gamma_I s_I'(i,c_i)$ and $s_W(j,b_j) = \gamma_W s_W'(j,b_j)$. Maximizing Eqn. (23) may get a byproduct, tag exemplars, and our approach mainly use them as a bridge to exploit tag information to help find image exemplars. We depict Eqn. (23) using a factor graph in Fig. 3. Each term in Eqn. (23) is represented by a *function node* and each label c_i (or b_i) is represented by a *variable node*. Edges exist only between function and variable nodes, and a function node is connected to a variable node iff its corresponding term depends on the variable. Heterogeneous relations e_{ij} serve as a bridge to connect two factor graphs over images and tags.

3.1 H^2 message propagation

This section presents our proposed *scalar* message propagation algorithm, homogeneous and heterogeneous message propagation (H^{2}MP), which transmits hybrid messages, over image and tag nodes, to maximize



(a) $r(i,k)$ (b) $a(i,k)$ (c) $w(i,j)$ (d) $v(i,j)$

Figure 5: Scalar-valued messages.

the objective function Eqn. (23). This algorithm starts from the *max-sum scheme*, and transform the *vector* massage propagation to *scalar* message propagation so that the algorithm is very fast.

Naive vector message propagation

We first present the naive vector message propagation algorithm. For simplicity, we only give the messages on the image side, and the messages on the tag side are similar. There are two vector messages between c_i and δ_k, and additionally another message from the heterogeneous relation node e_{ij}. The two vector messages are depicted in Figs. 4(a) and 4(b), and formulated as follows,

$$\rho_{i\to k}(c_i) = \sum_{e\in\mathcal{E}_i^R} \upsilon_{i\leftarrow e}(c_i) + s(i,c_i) + \sum_{k':k'\neq k} \alpha_{i\leftarrow k'}(c_i), \tag{12}$$

$$\begin{aligned}\alpha_{i\leftarrow k}(c_i) = & \max_{h_1,\cdots,h_{i-1},h_{i+1},\cdots,h_n} [\textstyle\sum_{i':i'\neq i} \rho_{i'\to k}(h_{i'}) \\ & + \delta_k(h_1,\cdots,h_{i-1},c_i,h_{i+1},\cdots,h_n)].\end{aligned} \tag{13}$$

Different from affinity propagation, we have additional two vector-valued messages exchanged between c_i and e_{ij}. The message, $\pi_{i\to e}$, sent from variable c_i to e_{ij}, consists of n real numbers, with one for each possible value of c_i. The message, $\upsilon_{i\leftarrow e}$, sent from variable e_{ij} to c_i, also consists of n real numbers. The two messages are depicted in Figs. 4(c) and 4(d), and formulated as follows,

$$\pi_{i\to e}(c_i) = s(i,c_i) + \sum_k \alpha_{i\leftarrow k}(c_i) + \sum_{e'\in\mathcal{E}_i^R/\{e\}} \upsilon_{i\leftarrow e'}(c_i), \tag{14}$$

$$\upsilon_{i\leftarrow e}(c_i) = \max_{b_j}[e(c_i,b_j) + \pi_{j\to e}(b_j)]. \tag{15}$$

One core of this paper is to reduce the above *vector-valued* messages to *scalar*-valued messages. The derivation is generalized from [8], but it is nontrivial and more challenging because our problem involves the heterogeneous relations that cannot be simplified directly using the derivation [8]. Due to space limitation, we omit the detail derivation[1] from vector messages to scalar messages. As a result, H^{2}MP views each image or tag as a node in a network, and recursively transmits *scalar-valued* messages along edges of the network until a good set of image and tag exemplars emerges. H^{2}MP is different from the original affinity propagation algorithm [8] in that H^{2}MP transmits not only the homogeneous messages within images and tags, including responsibility and availability, and depicted in Figs. 5(a) and 5(b), but also the heterogeneous messages between images and tags, including discardability and contributability, and depicted in Figs. 5(c) and 5(d). In the following, we will present four kinds of messages, and for convenience, we would only

[1] Please see the supplementary material if the reviewers are interested in the derivation.

Table 1: Comparison of the related methods on exploiting the homogeneous ("homo") and heterogeneous relations ("hetero"). "simultaneous" means that the two relations are simultaneously exploited, and "sequential" means that the two relations are sequentially exploited. Among these approaches, joint clustering [13] is very close to our approach, but exploits the heterogeneous relations in a sequential way.

	H^{2}MP	AP/k-means [17]	BGP	TGP	Joint [13]
Homo	Y	Y	N	partial	Y
Hetero	simultaneous	N	Y	Y	sequential

present the homogeneous messages over images as the messages over tags are similar and present the heterogeneous messages by standing at the image side as the messages on the tag side can also be similarly obtained. For presentation simplicity, we drop the subscripts I.

Homogeneous message propagation

The "responsibility" and "availability" messages in H^{2}MP are updated as follows,

$$r(i,k) = \bar{s}(i,k) - \max_{i':i'\neq k}[\bar{s}(i,i') + a(i,i')]. \tag{16}$$

$$\bar{s}(i,k) = \begin{cases} \sum_{j\in\mathcal{E}^R_{i\cdot}} v(i,j) + s(i,i) & k = i, \\ s(i,k) & k \neq i. \end{cases} \tag{17}$$

$$a(i,k) = \begin{cases} \sum_{i':i'\neq k} \max(0, r(i',k)) & k = i, \\ \min[0, r(k,k) + \sum_{i':i'\neq i,k} \max(0, r(i',k))] & k \neq i. \end{cases} \tag{18}$$

The key difference in the two messages from the original affinity propagation lies in the responsibility $r(i,j)$, which involves the heterogeneous message, i.e., the "contributability" message $v(i,j)$ from tag j to image i. This serves as a bridge in which the affect from tags will be transmitted to images. In the iteration process, $v(i,j)$ would become relatively larger when the probability of tag j being an exemplar becomes larger, and become smaller otherwise. Looking at Eqn. (37), we can observe that the contributability message takes effect when $k = i$, which means that it affects the preference of image i being an exemplar. Hence, the probability of image i, selecting itself as its exemplar, would be affected positively monotonically by the probability that tags linking to image i serve as exemplars.

Heterogeneous message propagation

There are two kinds of message exchanged between images and tags. The "discardability" $w(i,j)$, sent from image i to tag j, which reflects how much it is affected that image i selects itself as its exemplar when the contribution of word j is discarded and helps tag j make better decision whether to select itself as its exemplar. The "contributability" $v(i,j)$, sent from tag j to image i, which reflects how well image i serves as an exemplar considering whether tag j is an exemplar. The two messages are updated as

$$w(i,j) = r(i,i) + a(i,i) - v(i,j) = t(i,i) - v(i,j). \tag{19}$$

$$\begin{aligned} v(i,j) = &\max\{p(i,j), q(i,j) + w(j,i)\} \\ &- \max\{\bar{q}(i,j), p(j,i) + w(j,i)\}. \end{aligned} \tag{20}$$

Here, in Eqn. (39), $r(i,i)+a(i,i) = t(i,i)$ is the belief that image i selects itself as its exemplar, and $w(i,j)$ aims to evaluate the affect degree if the contribution from tag j to image i is discarded and help tag j make better decision whether to select itself as its exemplar. In evaluating the contributability message $v(i,j)$

from tag j to image i in Eqn. (40), $w(j,i)$ means that the belief that tag j selects itself as its exemplar without considering the contribution from image i, and $q(i,j)+w(j,i)$ evaluates the contribution from tag j to the probability that image i serves as an exemplar. $\max\{p(i,j), q(i,j)+w(j,i)\}$ essentially means that the degree that image i serves as an exemplar whether tag j serves as an exemplar. Similarly, $\max\{\bar{q}(i,j), p(j,i)+w(j,i)\}$ means that the degree that image i does not serve as an exemplar whether tag j serves as an exemplar. Their difference, called contributability, hence can evaluate how well image i serves as an exemplar considering the contribution from tag j. $v(i,j)>0$ means positive contribution from tag j, and negative contribution otherwise.

Exemplar assignment

The belief that image i selects image j as its exemplar is derived as the sum of the incoming messages,

$$t(i,j) = r(i,j) + a(i,j). \tag{21}$$

Then the exemplar of image i is taken as

$$\hat{c}_i = \arg\max\nolimits_{j \in \mathcal{E}^I_i \cup \{i\}} t(i,j). \tag{22}$$

It should be noted that the heterogeneous relations are latently involved in assigning the exemplars because the responsibility $r(i,j)$ already counts the contribution from tags that is indicated in Eqn. (36) and Eqn. (37).

To summarize, H^{2}MP is an iterative algorithm, and at the beginning all the eight kinds of messages are initialized as 0, and the eight messages are repeatedly updated until the iteration number reaches T or the identified exemplars do not change. The algorithm is described in the following,

Algorithm 1 Hybrid summarization

```
1. Initialize all the 8 messages as 0.
2. Compute 4 homogeneous messages for images and tags according to Eqn. (36) and Eqn. (38).
3. Compute 4 heterogeneous messages between images and tags according to Eqn. (39) and Eqn. (40).
4. Repeat steps 2 and 3 till the iteration number reaches $T$ or the identified exemplars do not change.
5. Make image exemplar assignments according to Eqn. (22).
```

3.2 Analysis and discussion

This subsection presents the time complexity analysis and discusses the relations of our approach with several existing approaches.

The naive implementation of H^{2}MP would take $O(T'(n^3+m^3+mn(m+n)))$ with T' the iteration number. Through the trick of reusing some computations, (e.g., $\max_{i':i'\neq i}[\bar{s}(i,i')+a(i,i')]$ in Eqn. (36), $\sum_{j\in\mathcal{E}^R_{i\cdot}} v(i,j)$ in Eqn. (37), and $\sum_{i'}\max[0, r(i',k)]$) in Eqn. (38), are just computed one time for each i in one iteration), the time complexity of our implementation is reduced to $O(T'(|\mathcal{E}^I|+|\mathcal{E}^W|+|\mathcal{E}^R|))$ with $|\cdot|$ being the edge number.

Our solution to hybrid image summarization is different from two previous representative techniques by Simon et al. [17] and Raguram and Lazebnik [13] in the following aspects. Simon et al. compute the visual summary by greedy k-means only using the visual information, without exploring the useful associated textual information. Raguram and Lazebnik use a joint clustering method, which first obtains two independent clusters from visual and textual features, respectively, and then takes their intersection to get the final clustering, but the joint process is obtained sequentially instead of simultaneously as our approach.

The proposed approach, H^{2}MP, is capable to exploit both the relations within images and tags and the relations between images and tags. Most related approaches are only able to capture partial relations. For example, AP (affinity propagation [8]) can only exploit homogeneous relations over images and words respectively, BGP (Bipartite graph partitioning [6]) can only exploit heterogeneous relations between images and tags, TGB (Tripartite graph partitioning [14]) uses the visual features besides heterogeneous relations between images and tags to help find the grouping without exploring the interior relations within tags. The comparison is summarized in Tab. 1.

3.3 Implementation

Similar to AP [8], the self-similarity $s'_I(i,i)$ of an image i, i.e., the preference of an image being an exemplar, is set as $\lambda\,\mathrm{Med}[s'_I(i,k)]$ with $\mathrm{Med}[s'_I(i,k)]$ being the median image similarity. λ is useful to control the exemplar number. For tags, we adopt the WordNet similarity [2], a variety of semantic similarity and relatedness measures based on a large lexical database of English, WordNet [3]. The self-similarities of words are similarly set.

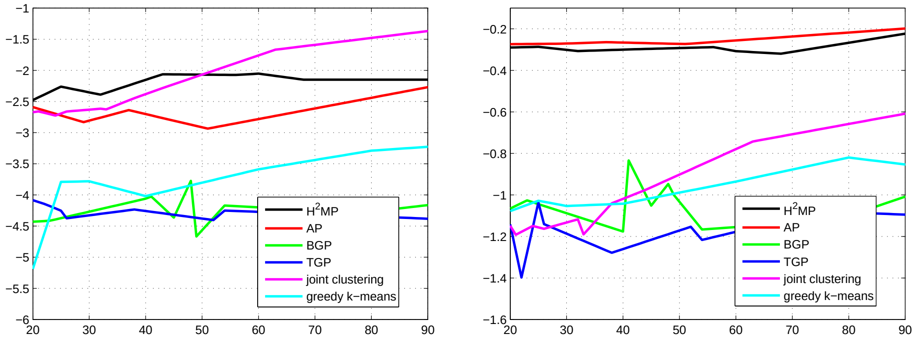
Let's turn to the setting of γ_I and γ_W in Eqn. (23). Looking at Eqn. (37), we observed that the heterogeneous relations essentially adjust the preference of image i, $\bar{s}(i,i)$, through the contributability $v(i,j)$ from tag j to i, and hence it is expected that $v(i,j)$ is comparable with the preference $s(i,i)$. Furthermore, we observed that $v(i,j)$ is computed from $w(j,i)$ in Eqn. (40) and $w(j,i)$ sent from tag j to image i is computed from the belief $t(j,j)$ of word j being an exemplar that is related to tag similarities in Eqn. (39). Thus, to make $p(\cdot,\cdot)$ in the heterogeneous relations easily tuned, which may benefit from the comparable preferences of tag and image, in our experiment we fix $\gamma_I = \frac{1}{\mathrm{Med}(s'_I(i,k))}$ and $\gamma_W = \frac{1}{\mathrm{Med}(s'_W(j,k))}$.

For $p(\cdot,\cdot)$ in the heterogeneous relations $e_{ij}(c_i,b_j)$, we set $p(i,j) = \theta/|\mathcal{E}^R_{i\cdot}|$, and $p(j,i) = \theta/|\mathcal{E}^R_{\cdot j}|$, where θ is a constant negative value, fixed as 15 in this paper, to control the mutual affect degree for image and tag exemplar identification, and the division by the tag number connecting image i, $|\mathcal{E}^R_{i\cdot}|$, aims to averagely separate its affect to connected tags.

4 Experiments

In our experiment, we present the performance comparison of our approach with several relate approaches. This collection consists of about 11k images and associated tags and is crawled from the popular photo sharing Web site Flickr, using the queries, including flower, city, building, dog, cat, plants, mountain, river, sunset, and so on. We filter out some noisy tags that few images are associated with and finally get 816 tags. On average, each image has 6.1 tags and each tag is assigned to 15.9 images. We extract a GIST scene descriptor [12], which has been shown to work well for scene categorization, as the image feature with 3 by 3 spatial resolution where each bin contains that image region's average response to steerable filters at 6 orientations and 3 scales, and use the negative Euclidean distance as the image similarity.

We investigate the performances from both the visual and semantic perspectives. In the literature of image summarization and clustering, most evaluation criteria use the class labels of the images to test the



(a)

(b)

Figure 6: Performance comparison with related approaches. The x-axis represents the exemplar number, the y-axis in (a) and (b) represent the semantic exemplarness and visual exemplarness.

performance. However, they are not adoptable for our hybrid summarization because hybrid summarization has multiple objectives, visual and sematic objectives and no simple labels can be applied here. Instead, we present two straightforward measures, visual exemplarness and sematic exemplarness. Visual exemplarness is defined as the average value of visual similarities between each image and its corresponding exemplar, and semantic exemplarness is defined as the average value of textual similarities between the associated tags of each image and its corresponding exemplar.

4.1 Quantitative comparison

We present a quantitative comparison of our approach (H^{2}MP) with several representative approaches, AP (affinity propagation [8]), BGP (Bipartite graph partitioning [6]), TGB (Tripartite graph partitioning [14]), and recently developed two methods: greedy k-means [17] and joint clustering [13]. Fig. 6(a) and Fig. 6(b) illustrate the performances of different approaches in terms of semantic and visual exemplarness with different number of exemplars.

For semantic exemplarness, H^{2}MP constantly outperforms the other approaches except the joint clustering approach [13] and its performance is a little worse than the joint clustering approach when the number of exemplars exceeds 50. This is understandable because our approach balances the visual and semantic performances while joint clustering generates results by taking intersection between the results using visual feature and text feature to cluster images, and hence may get superiority for semantic performance when the cluster number is very large. However, the performance for modest number of exemplars is more meaningful, because too many exemplars are not preferred in summarization. From this sense, our approach is more satisfactory in semantic performance.

For visual exemplarness, both AP and H^{2}MP show significant advantages over the other approaches. The visual performance of H^{2}MP is only a little worse than that of AP that purely uses visual feature, which is reasonable since our approach also takes into consideration the semantic information. In summary, H^{2}MP achieves satisfactory semantic and visual performance compared with other approaches.

Table 2: Quantitative results on three groups: “Anchorage, Alaska”, “Roma-Rome” and “The Great Wall of China”. “S” represents the semantic exemplarness, and “V” denotes the visual exemplarness. The images in “The Great Wall of China” have few tags and hence the semantic exemplarness measure can not be evaluated. The best two scores are highlighted in bold font.

	S	V	S	V	S	V
H^2MP	**-2.264**	-0.761	**-0.901**	-0.775	-	**-0.399**
AP	-3.225	**-0.745**	-2.473	**-0.706**	-	**-0.396**
BGP	-4.161	-0.938	-4.637	-0.919	-	-0.407
TGP	-3.702	-0.928	-3.266	-0.802	-	-0.593
Joint [13]	**-2.205**	-0.917	-3.596	-1.062	-	- 0.499
k-means [17]	-2.283	**-0.703**	**-2.202**	**-0.650**	-	-0.412

4.2 Visual comparison

We present visual results on three representative groups of images from Flickr, “Anchorage, Alaska”, “Roma-Rome”, and “The Great Wall of China”. We crawled top 970, 928, and 133 images, respectively.

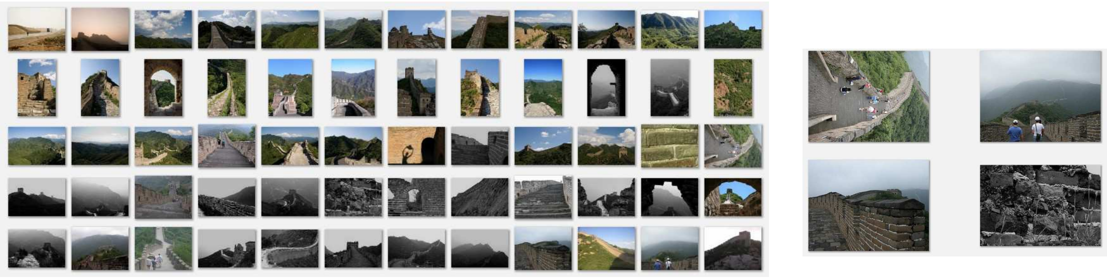
The visual results of “Anchorage, Alaska” “Roma-Rome” from the six approaches are depicted in Fig. 8, and their quantitative comparison is in Tab. 2. Our results look visually appealing, and the superiority in semantic performances shows that the obtained visual summary can capture the semantic meaning, which benefits from the associated tags. The other methods cannot get competitive performance because those methods have partial or little ability to exploit homogeneous and heterogeneous relations.

4.3 User study

In addition, we present a user study to compare visual summaries of six approaches. We collect the feedback from 20 persons on 10 tagged image collections. For each person, we show him a tagged image collection and randomly select a visual summary from the six results corresponding six approaches, and allow him to given a score from 1 (the worst) to 5 (the best) to indicate how well the visual summary represents the image collection from the visual and semantic perspectives. The user study shows that our approach obtains the best performance 4.3, and the scores of other approaches, AP, BGP, TGP, joint clustering [13] and greedy k-means [17] are 3.6, 3, 3.2, 3.9, and 3.8, respectively. This user study demonstrates that our approach can get better visual summary compared with other approaches.

4.4 Application

We demonstrate the hybrid summarization in Flickr group overview by presenting both image and text summarization. “Flickr groups are a fabulous way to share content and conversation, either privately or with the world. Believe us when we say there’s probably a group for everyone, but if you can’t find one you like, feel free to start your own.”. The group images are displayed page by page, and each page shows a dozen of images. To have an overview of a group of images, uses have to check the images page by page, and also there is no textual description for the group of images. Hence it is desired to deliver visual and textual summaries of the group. Fig. 7 shows an example over “The Great Wall of China”. It is surprisedly that the hybrid summarization suggests two tags: simatai and mutianyu. After checking this group manually,



(b) Visual summary

simatai, mutianyu

(a) Sample images in the group

(c) Textual summary

Figure 7: Flickr group summarization for “The Great Wall of China”.

we found that this group only contains the photos from two sites.

5 Conclusion

In this paper, we present hybrid image summarization scheme to manage image collections. Toward this end, we propose a novel approach, homogeneous and heterogeneous message propagation, which is a non-trivial generalization of the affinity propagation algorithm from homogeneous data to heterogeneous data. Compared with the conventional message propagation algorithms that transmit the vector-valued messages, our algorithm reduces vector-valued messages to scalar-valued messages, and hence is more efficient. Moreover, this reduction in our case is more complicated than in affinity propagation because it involves additional heterogeneous relations. The application of our approach to hybrid image summarization can effectively exploit image similarities and even the useful information from the associated tags, including the association relations between images and tags and the relations within tags. The experimental results demonstrate its effectiveness and efficiency.

Appendix

Derivation

We rewrite the objective function, which corresponds to Eqn. (11) in the submitted paper.

$$\begin{aligned}\mathcal{S}(\mathbf{c},\mathbf{b}) =& \gamma_I(E^I(\mathbf{c})+V^I(\mathbf{c}))+\gamma_W(E^W(\mathbf{b})+V^W(\mathbf{b}))+R(\mathbf{c},\mathbf{b})\\ =& \sum\nolimits_{i=1}^{n} s_I(i,c_i)+\sum\nolimits_{k=1}^{n}\delta_k(\mathbf{c})+\sum\nolimits_{j=1}^{m} s_W(j,b_j)+\sum\nolimits_{k=1}^{m}\eta_k(\mathbf{b})+\sum\nolimits_{(i,j)\in\mathcal{E}^R} e_{ij}(c_i,b_j).\end{aligned} \tag{23}$$

This objective function is depicted as a factor graph Fig. 3. The max-sum algorithm [11], a general algorithm to factor graph, can get the solution by transmitting vector messages between function nodes and variable nodes. We would derive a scalar message propagation scheme, homogeneous and heterogeneous message propagation, by reducing vector messages over function and variable nodes to scalar messages over variable nodes.

The max-sum algorithm is an iterative algorithm to exchange two kinds of messages: one is from function nodes to variable nodes, and the other is from variable nodes to function nodes. For the factor graph Fig. 3 corresponding to Eqn. (23), the message propagation over variables c_i and b_j is almost the same, For convenience we will give the derivation over variable c_i, and drop the subscript I in $s_I(\cdot,\cdot)$ in the following presentation.

There are two messages exchanged between c_i and δ_k. The message, $\rho_{i\to k}$, sent from c_i to δ_k, consists of n real numbers, one for each possible value of c_i. The message, $\alpha_{i\leftarrow k}$, sent from δ_k to c_i, also consists of n real numbers. The two messages are depicted in Figs. 4(a) and 4(b), and formulated as follows:

$$\rho_{i\to k}(c_i) = \sum_{e\in\mathcal{E}_i^R} \upsilon_{i\leftarrow e}(c_i) + s(i,c_i) + \sum_{k':k'\neq k} \alpha_{i\leftarrow k'}(c_i), \tag{24}$$

$$\alpha_{i\leftarrow k}(c_i) = \max_{h_1,\cdots,h_{i-1},h_{i+1},\cdots,h_n} [\textstyle\sum_{i':i'\neq i} \rho_{i'\to k}(h_{i'}) + \delta_k(h_1,\cdots,h_{i-1},c_i,h_{i+1},\cdots,h_n)]. \tag{25}$$

There are two messages exchanged between c_i and e_{ij}. The message, $\pi_{i\to e}$, sent from variable c_i to e_{ij}, consists of n real numbers, one for each possible value of c_i. The message, $\upsilon_{i\leftarrow e}$, sent from variable e_{ij} to c_i, also consists of n real numbers. Let $\mathcal{E}_i^R = \{e_{ij}\}_j$ represent the edge set connecting image i. For simplicity, we drop the subscript $_ij$ in e_{ij} without influencing understanding. The two messages are depicted in Figs. 4(c) and 4(d), and formulated as follows:

$$\pi_{i\to e}(c_i) = s(i,c_i) + \sum_k \alpha_{i\leftarrow k}(c_i) + \sum_{e'\in\mathcal{E}_i^R/\{e\}} \upsilon_{i\leftarrow e'}(c_i), \tag{26}$$

$$\upsilon_{i\leftarrow e}(c_i) = \max_{b_j}[e(c_i,b_j) + \pi_{j\to e}(b_j)]. \tag{27}$$

In the following, we show that those vector-valued messages can be reduced to scalar-valued messages, making the propagation much more efficient. The derivation is generalized from [8], but it is nontrivial and more challenging since the message is additionally propagated between heterogeneous data, images and words. We directly present the results for $\rho-$ and $\alpha-$ messages by omitting detailed derivation that can be obtained using the similar technique as in [8]. We present the derivation detail for $\upsilon-$ and $\pi-$ messages. The idea behind the derivation is to analyze the propagated messages in the two cases whether c_i is valued as i or not.

Let $\tilde{\rho}_{i\to k}(c_i) = \rho_{i\to k}(c_i) - \bar{\rho}_{i\to k}$, with $\bar{\rho}_{i\to k} = \max_{h:h\neq k} \rho_{i\to k}(h)$.

Let $\tilde{\alpha}_{i\leftarrow k}(c_i) = \alpha_{i\leftarrow k}(c_i) - \bar{\alpha}_{i\leftarrow k}$, with $\bar{\alpha}_{i\leftarrow k} = \alpha_{i\leftarrow k}(c_i : c_i \neq k)$. It can be derived that $\alpha_{i\leftarrow k}(c_i : c_i \neq k)$ is independent to the specific value c_i.

Let $\tilde{\upsilon}_{i\leftarrow e}(c_i) = \upsilon_{i\leftarrow e}(c_i) - \bar{\upsilon}_{i\leftarrow e}$,

$$\begin{aligned}\bar{\upsilon}_{i\leftarrow e} &= \upsilon_{i\leftarrow e}(c_i : c_i \neq i)\\ &= \max_{b_j}[e(c_i,b_j) + \pi_{j\to e}(b_j)]\\ &= \max[\max_{b_j:b_j\neq j}[e(c_i,b_j) + \pi_{j\to e}(b_j)], e(c_i,j) + \pi_{j\to e}(j)]\\ &= \max[\bar{q}(i,j) + \max_{b_j:b_j\neq j} \pi_{j\to e}(b_j), p(j,i) + \pi_{j\to e}(j)].\end{aligned} \tag{28}$$

Let $\tilde{\pi}_{i\to e}(c_i) = \pi_{i\to e}(c_i) - \bar{\pi}_{i\to e}(c_i)$, and

$$\begin{aligned}\bar{\pi}_{i\to e} &= \max\nolimits_{c_i:c_i\neq i} \pi_{i\to e}(c_i)\\ &= \max_{c_i:c_i\neq i}[s(i,c_i) + \tilde{\alpha}_{i\leftarrow c_i}(c_i) + \sum_{e'\in\mathcal{E}_i^R/\{e\}} \upsilon_{i\leftarrow e'}(c_i) + \sum_k \bar{\alpha}_{i\leftarrow k}]\\ &= \max_{c_i:c_i\neq i}[s(i,c_i) + \tilde{\alpha}_{i\leftarrow c_i}(c_i)] + \sum_{e'\in\mathcal{E}_i^R/\{e\}} \upsilon_{i\leftarrow e'}(c_i\neq i) + \sum_k \bar{\alpha}_{i\leftarrow k}. \qquad (29)\end{aligned}$$

For $\tilde{\rho}_{i\to k}(c_i = k)$ and $\tilde{\alpha}_{i\leftarrow k}(c_i = k)$, we can obtain

$$\tilde{\rho}_{i\to k}(c_i = k) = \bar{s}(i,k) - \max\nolimits_{i':i'\neq k}[\bar{s}(i,i') + \alpha_{i\leftarrow k}(c_i = i')]. \qquad (30)$$

$$\bar{s}(i,k) = \begin{cases} \sum_{e\in\mathcal{E}_i} \upsilon_{i\leftarrow e}(c_i = i) + s(i,i), & k = i\\ s(i,k), & k\neq i. \end{cases} \qquad (31)$$

$$\tilde{\alpha}_{i\leftarrow k}(c_i = k) = \begin{cases} \sum_{i':i'\neq k} \max(0, \tilde{\rho}_{i'\to k}(c_i = k)), & k = i\\ \min[0, \tilde{\rho}_{k\to k}(c_k = k)) + \sum_{i':i'\neq i,k} \tilde{\rho}_{i'\to k}(c_i = k))], & k\neq i. \end{cases} \qquad (32)$$

For $\tilde{\upsilon}$ and $\tilde{\pi}$, we have the following derivations

$$\begin{aligned}&\tilde{\upsilon}_{i\leftarrow e}(c_i = i)\\ =&\upsilon_{i\leftarrow e}(c_i = i) - \bar{\upsilon}_{i\leftarrow e}\\ =&\max[p(i,j) + \max\nolimits_{b_j:b_j\neq j} \pi_{j\to e}(b_j), q(i,j) + \pi_{j\to e}(j)] - \max[\bar{q}(i,j) + \max\nolimits_{b_j:b_j\neq j} \pi_{j\to e}(b_j), p(j,i) + \pi_{j\to e}(j)]\\ =&\max[p(i,j), q(i,j) + \pi_{j\to e}(j) - \max\nolimits_{b_j:b_j\neq j} \pi_{j\to e}(b_j)] - \max[\bar{q}(i,j)+, p(j,i) + \pi_{j\to e}(j) - \max\nolimits_{b_j:b_j\neq j} \pi_{j\to e}(b_j)]\\ =&\max[p(i,j), q(i,j) + \tilde{\pi}_{j\to e}(j)] - \max[\bar{q}(i,j), p(j,i) + \tilde{\pi}_{j\to e}(j)]. \qquad (33)\end{aligned}$$

$$\tilde{\upsilon}_{i\leftarrow e}(c_i\neq i) = \upsilon_{i\leftarrow e}(c_i\neq i) - \bar{\upsilon}_{i\leftarrow e} = 0. \qquad (34)$$

$$\begin{aligned}&\tilde{\pi}_{i\to e}(c_i = i)\\ =&\pi_{i\to e}(c_i = i) - \bar{\pi}_{i\to e}\\ =&s(i,i) + \tilde{\alpha}_{i\leftarrow i}(i) - \max_{c_i:c_i\neq i}[s(i,c_i) + \tilde{\alpha}_{i\leftarrow c_i}(c_i)] + \sum_{e'\in\mathcal{E}_i/\{e\}} \tilde{\upsilon}_{i\leftarrow e'}(i). \qquad (35)\end{aligned}$$

It can observed that only the variables $\tilde{\rho}_{i\to k}(c_i)$ and $\tilde{\alpha}_{i\leftarrow k}(c_i)$ for $c_i = k$ and $\tilde{\upsilon}_{i\leftarrow e}(c_i)$ and $\tilde{\pi}_{i\to e}(c_i)$ for $c_i = i$ are involved in the message passing. Therefore, we can define scalar-valued variables $r(i,k) = \tilde{\rho}_{i\to k}(c_i = k)$, $a(i,k) = \tilde{\alpha}_{i\leftarrow k}(c_i = k)$, $v(i,j) = \tilde{\upsilon}_{i\leftarrow e_{ij}}(c_i = i)$, and $w(i,j) = \tilde{\pi}_{i\to e_{ij}}(c_i = i)$. These scalar messages are summarized as follows.

Homogeneous message propagation

There are two kinds of messages exchanged within image points. The "responsibility" $r(i,k)$, sent from data point i to data point k, which reflects how well k serves as the exemplar of i considering other potential exemplars for i, and the "availability" $a(i,k)$, sent from data point k to data point i, which reflects how appropriately i chooses k as its exemplar considering other potential points that may choose k as their exemplar. The messages are updated in an iterative way as

$$r(i,k) = \bar{s}(i,k) - \max_{i':i'\neq k}[\bar{s}(i,i') + a(i,i')]. \tag{36}$$

$$\bar{s}(i,k) = \begin{cases} \sum_{j\in\mathcal{E}^R_{i\cdot}} v(i,j) + s(i,i) & k = i, \\ s(i,k) & k \neq i. \end{cases} \tag{37}$$

$$a(i,k) = \begin{cases} \sum_{i':i'\neq k} \max(0, r(i',k)) & k = i, \\ \min[0, r(k,k) + \sum_{i':i'\neq i,k} \max(0, r(i',k))] & k \neq i. \end{cases} \tag{38}$$

Heterogeneous message propagation

There are two kinds of message exchanged between images and words. The "discardability" $w(i,j)$, sent from image i to word j, which reflects how much it is affected that image i selects itself as its exemplar when the contribution of word j is discarded and helps word j make better decision whether to select itself as its exemplar. The "contributability" $v(i,j)$, sent from word j to image i, which reflects how well image i serves as an exemplar considering whether word j is an exemplar. The two messages are updated as

$$w(i,j) = r(i,i) + a(i,i) - v(i,j) = t(i,i) - v(i,j). \tag{39}$$

$$\begin{aligned} v(i,j) = &\max\{p(i,j), q(i,j) + w(j,i)\} \\ &- \max\{\bar{q}(i,j), p(j,i) + w(j,i)\}. \end{aligned} \tag{40}$$

Exemplar assignment

To obtain exemplar assignment after convergence, we sum together all the incoming messages to c_i and take the value $\hat{c}_i$ as follows:

$$\begin{aligned} \hat{c}_i = &\arg\max_{i'}[\textstyle\sum_{i'} \alpha_{i\leftarrow k}(i') + s(i,i') + \sum_j \upsilon_{i\leftarrow e_{ij}}(i')] \\ = &\arg\max_{i'}[\textstyle\sum_k \tilde{\alpha}_{i\leftarrow k}(i') + \sum_k \bar{\alpha}_{i\leftarrow k} + s(i,i') + \sum_j \tilde{\upsilon}_{i\leftarrow e_{ij}}(i') + \sum_j \bar{\upsilon}_{i\leftarrow e_{ij}}] \\ = &\arg\max_{i'} \begin{cases} a(i,i') + s(i,i') & i' \neq i \\ a(i,i') + \sum_j v(i,j) + s(i,i') & i' = i \end{cases} \\ = &\arg\max_{i'}[a(i,i') + \bar{s}(i,i')] \\ = &\arg\max_{i'}[a(i,i') + \bar{s}(i,i') - \max_{k:k\neq i'}(a(i,k) + \bar{s}(i,k)] \\ = &\arg\max_{i'}[a(i,i') + r(i,i')]. \end{aligned} \tag{41}$$

Complexity analysis

This section presents the complexity analysis of the proposed algorithm. The naive implementation of heterogeneous affinity propagation would take $O(n^3 + m^3 + mn(m+n))$ per iteration. In the following, we analyze the algorithm carefully and justify the algorithm essentially only costs $O(|\mathcal{E}^I| + |\mathcal{E}^W| + |\mathcal{E}^R|)$

per iteration through the trick of reusing some computations. The analysis borrows some ideas from [8], but differs from it because in our algorithm the responsibility message involves the sum of contributability messages and heterogeneous message propagation is additionally introduced.

When computing the responsibility message in Eqn. (36), the maximum and next-to-maximum values of $\bar{s}(i,i') + a(i,i')$ w.r.t. i' are computed one time for each i over one pass of the whole algorithm. Then the maximum value $\max_{i':i'\neq k}[\bar{s}(i,i') + a(i,i')]$ needed in Eqn. (36), can be found in a single operation, by checking to see if k gives the maximum (in which case the next-to-maximum value is used) or not (in which case the maximum value is used). When computing $\bar{s}(i,k)$, the summation of $v(i,j)$ w.r.t j is computed one time for each i over one pass of the whole algorithm. Similar tricks can also be used to evaluate $\sum_{i'} \max[0, r(i',k)])$ for computing $a(i,k)$.

All the messages are transmitted over the edges, and hence there are totally $O(|\mathcal{E}^I| + |\mathcal{E}^W| + |\mathcal{E}^R|)$ messages. The reused computations, the maximum and next-to-maximum values of $\bar{s}(i,i')+a(i,i')$, and the summation of $v(i,j)$ w.r.t j, are performed one time for each i, which cost $O(|\mathcal{E}_i^T|)$ and $O(|\mathcal{E}_{i.}^R|)$. Thence the reused computations cost $O(\sum_i |\mathcal{E}_i^I| + \sum_j |\mathcal{E}_j^W| + \sum_i |\mathcal{E}_{i.}^R| + \sum_j |\mathcal{E}_{.j}^R|) = O(|\mathcal{E}^I| + |\mathcal{E}^W| + |\mathcal{E}^R|)$. The exemplar assignment for images and words will cost $\sum_i |\mathcal{E}_i^I| + \sum_j |\mathcal{E}_j^W|$. In summary, the proposed algorithm costs $O(T'(|\mathcal{E}^I| + |\mathcal{E}^W| + |\mathcal{E}^R|))$ with T' being the iteration number.

References

[1] `http://picasaweb.google.com/`.

[2] `http://search.cpan.org/dist/WordNet-Similarity`.

[3] `http://wordnet.princeton.edu/`.

[4] `http://www.flickr.com/groups/`.

[5] R. Clough, H. Joho, and M. Sanderson. Automatically organising images using concept hierarchies. In *SIGIR Workshop on Multimedia Information Retrieval*, 2005.

[6] I. S. Dhillon. Co-Clustering Documents and Words using Bipartite Spectral Graph Partitioning. In *KDD*, pages 269–274, 2001.

[7] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-Theoretic Co-Clustering. In *KDD*, pages 89–98, 2003.

[8] B. J. Frey and D. Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315:972–976, February 2007.

[9] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q. Cheng, and W.-Y. Ma. Web Image Clustering by Consistent Utilization of Visual Features and Surrounding texts. In *ACM Multimedia*, pages 112–121, 2005.

[10] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating Summaries for Large Collections of Geo-Referenced Photographs. In *WWW*, pages 853–854, 2006.

[11] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor Graphs and the Sum-product Algorithm. *IEEE Trans. Information Theory*, 47(2):498–519, 2001.

[12] A. Oliva and A. B. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[13] R. Raguram and S. Lazebnik. Computing Iconic Summaries for General Visual Concepts. In *First IEEE Workshop on Internet Vision*, 2008.

[14] M. Rege, M. Dong, and J. Hua. Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering. In *WWW*, pages 317–326, 2008.

[15] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake. Digital Tapestry. In *CVPR (1)*, pages 589–596, 2005.

[16] P. Schmitz. Inducing Ontology from Flickr Tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, 2006.

[17] I. Simon, N. Snavely, and S. M. Seitz. Scene Summarization for Online Image Collections. In *ICCV*, 2007.

[18] J. Wang, J. Sun, L. Quan, X. Tang, and H.-Y. Shum. Picture Collage. In *CVPR (1)*, pages 347–354, 2006.

[19] H. Xu, J. Wang, X.-S. Hua, and S. Li. Hybrid image summarization. In *ACM Multimedia*, pages 1217–1220, 2011.

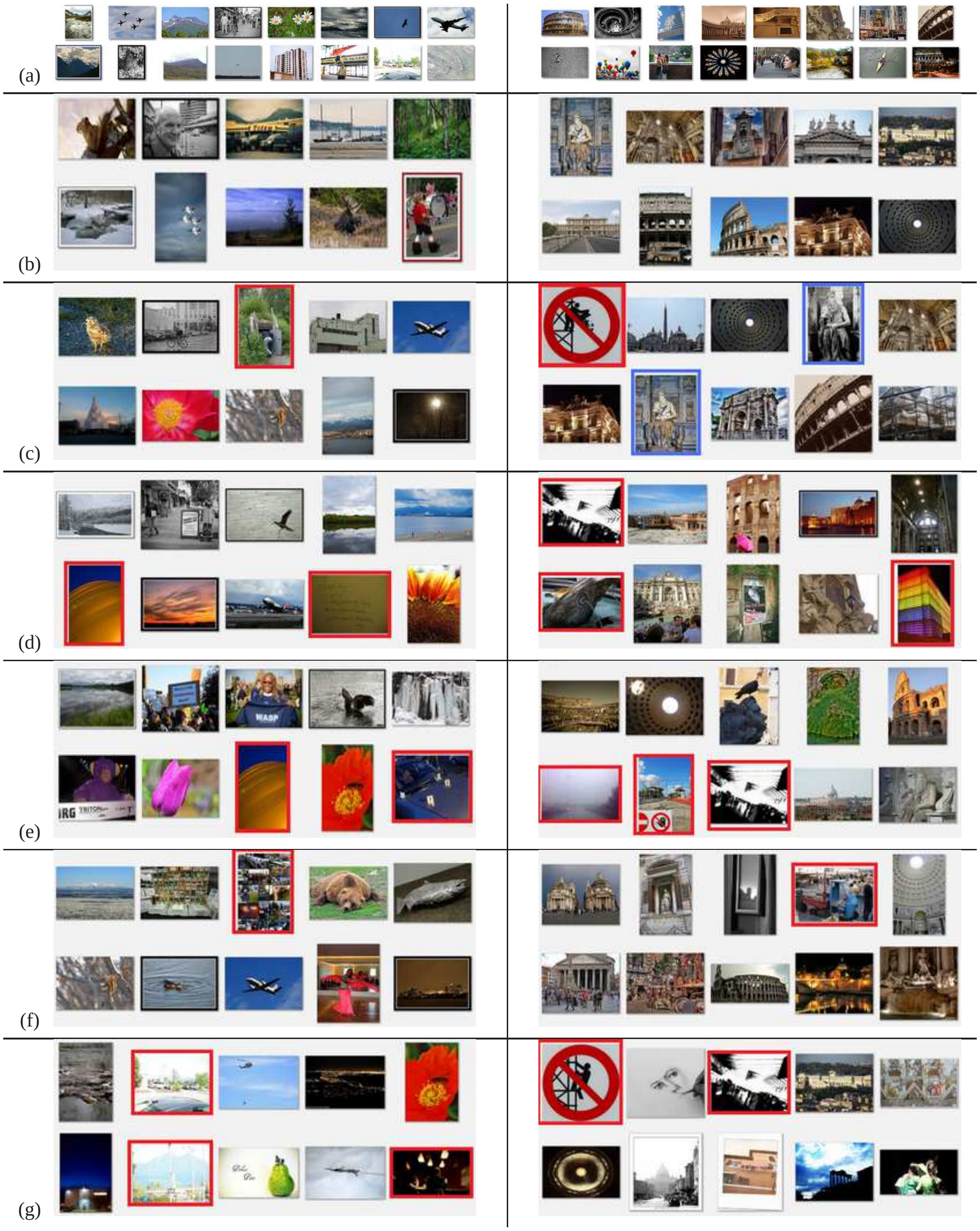


Figure 8: Visual results on “Anchorage, Alaska” (left column) and “Roma-Rome” (right column). (a) the examples of original images, (b)- (g) correspond to the results of H^{2}MP, AP, BGP, TGP, joint clustering and greedy k-means.