# Hybrid Image Summarization[*] [†]

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

## 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 ($H^2MP$), which extends affinity propagation that only works over homogeneous relations to heterogeneous relations. The summary obtained by our approach is both visually and semantically satisfactory. The experimental results demonstrate the effectiveness and efficiency of the proposed approach.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

## General Terms

Algorithms, Experimentation

## Keywords

Image summarization, image presentation, affinity propagation, homogeneous and heterogeneous message propagation

## 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, 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. [16] 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. [20, 12] 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. Wang et al. [19] present a hierarchial clustering scheme that can benefit the browsing of large image collections.

Recently, there are a few works to deal with the selection problem. Jia et al. [11] present a hierarchical affinity propagation approach to image collection summarization. Xu et al. [22] propose to describe images using cross-media information and perform an affinity propagation based image clustering approach. Simon et al. [18] 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 [14] 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

[†]Area Chair: Marcel Worring.
[*]This work was performed at Microsoft Research Asia.


*MM* '11 Arizona, USA

(a) (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.**

the final clustering, but the joint process is obtained sequentially instead of simultaneously. Surrounding texts are limitedly exploited for image grouping [9, 15] 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 [17] 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. 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.

# 2. IMAGE SUMMARIZATION

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.

## 2.1 H$^2$ message propagation

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. 3(a) and 3(b), but also the heterogeneous messages between images and tags, including discardability and contributability, and depicted in Figs. 3(c) and 3(d). In the following, we will present four kinds of messages, and for convenience, we would only 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,

$$r(i,k) = \bar{s}(i,k) - \max_{i':i' \neq k}[\bar{s}(i,i') + a(i,i')]. \quad (1)$$

$$\bar{s}(i,k) = \begin{cases} \sum_{j \in \mathcal{E}_i^R} v(i,j) + s(i,i) & k = i, \\ s(i,k) & k \neq i. \end{cases} \quad (2)$$

$$a(i,k) = \begin{cases} \sum_{i':i' \neq k} \max(0, r(i',k)) & k = i, \\ \min[0, r(k,k) + \sum_{i':i' \neq i,k} \max(0, r(i',k))] & k \neq i. \end{cases} \quad (3)$$

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. (2), 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 posi-

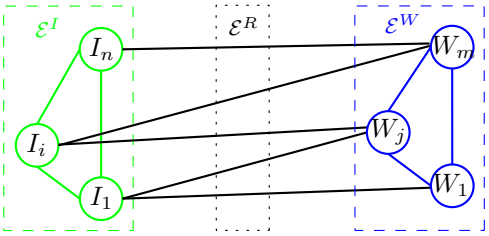
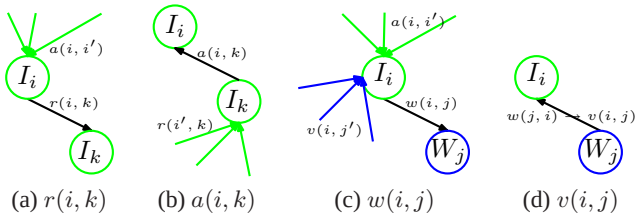



**Figure 2: Heterogeneous graph over images and tags.**




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

**Figure 3: Illustration of message propagation.**

tively 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

$$\begin{aligned} w(i,j) &= r(i,i) + a(i,i) - v(i,j) \\ &= t(i,i) - v(i,j). \end{aligned} \tag{4}$$

$$\begin{aligned} v(i,j) = &\max\{p(i,j), q(i,j) + w(j,i)\} \\ &- \max\{\bar{q}(i,j), p(j,i) + w(j,i)\}. \end{aligned} \tag{5}$$

Here, in Eqn. (4), $r(i,i) + a(i,i) = t(i,i)$ 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. (5), $w(j,i)$ 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,

$$t(i,j) = r(i,j) + a(i,j). \tag{6}$$

Then the exemplar of image $i$ is taken as

$$\hat{c}_i = \arg\max_{j \in \mathcal{E}_i^I \cup \{i\}} t(i,j). \tag{7}$$

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. (1) and Eqn. (2).

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. (1) and Eqn. (3).
3. Compute 4 heterogeneous messages between images and tags according to Eqn. (4) and Eqn. (5).
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. (7).

## 2.2 Implementation

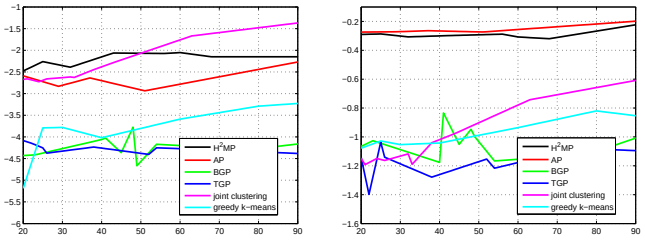
Similar to AP [8], the self-similarity $s'_I(i,i)$ of an image $i$, i.e., the preference of an image being an exemplar, is set as $\lambda \,\mathrm{Med}[s'_I(i,k)]$ with $\mathrm{Med}[s'_I(i,k)]$ being the median image similarity. $\lambda$ is useful to control the exemplar number. For tags, we adopt 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 consider $p(i,j)$ in the heterogeneous message. $p(i,j)$ is the weight describing the case $c_i = i, b_j \neq j$, and $p(j,i)$ is the weight describing the case $c_i \neq i, b_j = j$. We set $p(i,j) = \theta/|\mathcal{E}_{i\cdot}^R|$, and $p(j,i) = \theta/|\mathcal{E}_{\cdot j}^R|$, where $\theta$ 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 $i$, $|\mathcal{E}_{i\cdot}^R|$, aims to averagely separate its affect to connected tags. $q(i,j)$ and $\bar{q}(i,j)$ are the weights corresponding to the case $c_i \neq i, b_j \neq j$ and $c_i = i, b_j = j$.

# 3. 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 [13], 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 performance. However, they are not adoptable for our hybrid summarization because hybrid summarization has multiple objectives, visual and sematic objectives and no simple la-




(a) (b)

**Figure 4: Performance comparison. The x-axis represents the exemplar number, the y-axis in (a) and (b) represent the semantic exemplarness and visual exemplarness.**

bels 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.

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 [15]), and recently developed two methods: greedy k-means [18] and joint clustering [14]. Fig. 4(a) and Fig. 4(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 [14] 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.

## 4. CONCLUSION

In this paper, we present a hybrid image summarization scheme to manage image collections. Toward this end, we propose a novel approach, homogeneous and heterogeneous message propagation, which extends the affinity propagation algorithm from homogeneous data to heterogeneous data. The reduction from vector-valued messages to scalar-valued messages is more complicated than in affinity propagation because it involves additional heterogeneous relations and details can be found from [21]. The application of our approach to hybrid image summarization can effectively exploit image similarities, the association relations between images and tags and the relations within tags. The experimental results demonstrate its effectiveness and efficiency.

## 5. 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] Y. Jia, J. Wang, C. Zhang, and X.-S. Hua. Finding image exemplars using fast sparse affinity propagation. In *ACM Multimedia*, pages 639–642, 2008.

[12] T. Liu, J. Wang, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum. Picture collage. *IEEE Transactions on Multimedia*, 11(7):1225–1239, 2009.

[13] 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.

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

[15] 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.

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

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

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

[19] J. Wang, L. Jia, and X.-S. Hua. Interactive browsing via diversified visual summarization for image search results. *Multimedia Systems*.

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

[21] J. Wang, H. Xu, X.-S. Hua, and S. Li. Hybrid image summarization. Technical report, MSR-TR-2010-176, 2010.

[22] H. Xu, J. Wang, X.-S. Hua, and S. Li. Summarizing tagged image collections by cross-media representativeness voting. In *ICME*, pages 922–925, 2009.