A Joint Optimization Model for Image Summarization Based on Image Content and Tags

Hongliang Yu [†], **Zhi-Hong Deng** [†*], **Yunlun Yang** [†], and **Tao Xiong** [‡]
[†]Key Laboratory of Machine Perception (Ministry of Education),
School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[‡]Department of Electrical and Computer Engineering, The Johns Hopkins University
yuhongliang324@gmail.com, zhdeng@cis.pku.edu.cn, incomparable-lun@pku.edu.cn, tao.xiong@jhu.edu

Abstract

As an effective technology for navigating a large number of images, image summarization is becoming a promising task with the rapid development of image sharing sites and social networks. Most existing summarization approaches use the visual-based features for image representation without considering tag information. In this paper, we propose a novel framework, named JOINT, which employs both image content and tag information to summarize images. Our model generates the summary images which can best reconstruct the original collection. Based on the assumption that an image with representative content should also have typical tags, we introduce a similarity-inducing regularizer to our model. Furthermore, we impose the lasso penalty on the objective function to yield a concise summary set. Extensive experiments demonstrate our model outperforms the state-of-the-art approaches.

Introduction

With the rapid development of image sharing sites and social networks, it is difficult for users to search what they are interested in from a large amount of images in Internet. Image summarization, which aims to select a small set of representative images from a large-scale collection, has become a promising task. Various multimedia applications can benefit from image summarization. Wang, Jia, and Hua (2011) improve the image search results by a visual summarization technique. Yang et al. (2012) think travel websites should choose the most representative photos of tourist sites from the large-scale gallery when displaying on their web pages.

Most existing summarization approaches use the visual-based features for image representation, such as color histograms (Hafner et al. 1995), SIFT (Lowe 1999) or Bag-of-Visual-Words (Yang et al. 2007). Besides, in real world applications, many images have user-annotated tags, which provide additional information for image understanding. However, most prior work fails to utilize these text tags in the summarization process. Our work is motivated by the subject of how the image tags can be leveraged for image summarization rather than image content only.

*Corresponding author

Tags: *Greenland*, *Uummannaq*, *travel*, *icebergs*, *fjord*, *houses*, *summer*, *I love Greenland*

Figure 1: A landscape picture and its tags. The underlined tags cannot be directly found in the picture.

There are three advantages by using image tags. Firstly, compared with the visual content of an image, we are able to conveniently recognize the objects in the picture from the tags in a more explicit way. Secondly, users prefer to annotate the objects that reflect the topic of the image and ignore the unimportant elements. Figure 1 shows a landscape image and its tags. The user annotates "*icebergs*", "*fjord*" and "*houses*", but no tags like "*mountains*" or "*sky*" which also appears in the picture. This indicates the objects like "*mountains*" or "*sky*" are considered trivial. Thirdly, some tags imply invisible but highly relevant elements of images. We gain additional information from these tags, and can build relationships to other images expressing the same concept. In Figure 1, from the underlined tags, some hidden information can be found, e.g. "*Greenland*" and "*Uummannaq*" as the location and "*summer*" as the season.

In this paper, we propose a novel framework based on sparse coding for automatic image summarization, called JOINT (abbreviation for "a **J**oint **O**ptimization model for **I**mage summarization based on image Co**n**tent and **T**ags"). The distinctive aspect of our approach is that JOINT adopts both the image content and the corresponding tags in order to narrow the semantic gap. To effectively leverage available image tags, we argue that a good summary should consist of representative images whose tags are also representative. In our framework, the reconstruction error is defined to measure the "representativeness" of the images and tags. Since

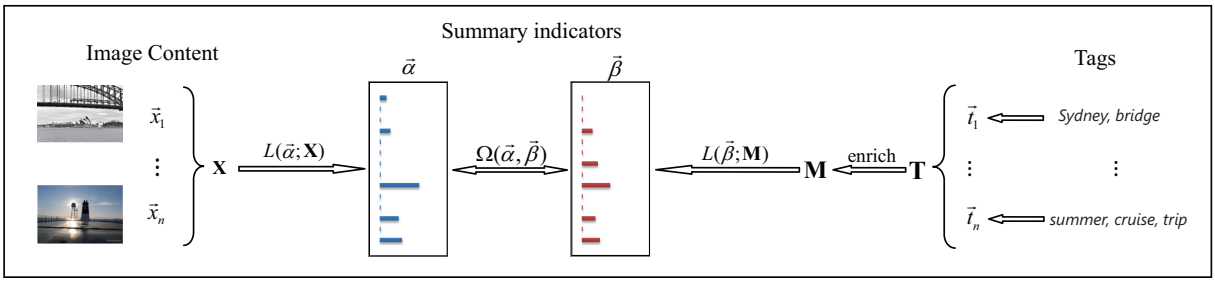

Figure 2: Schematic illustration.

several images lack of tags, we introduce the technology of matrix completion (Cai, Candès, and Shen 2010) to recover the missing tags.

We study the problem of image summarization and formulate it as an optimization problem. In our model, two indicator vectors, whose non-zero elements represent the summary images from the visual and textual viewpoints respectively, control which images are selected. We assume the images with representative image content should also have typical tags. Accordingly, a similarity-inducing regularizer imposed on the indicators is introduced to our model. We discuss two types of the regularizer in this paper. Furthermore, as the summary set is a small subset of the original image collection, we expect the indicator vectors to be sparse. To address the issue, we introduce the lasso penalty (Tibshirani 1996) to yield sparse solution. Finally, the images whose corresponding elements are non-zero in both indicators are selected as the summary.

To sum up, our contributions are:

(1) We propose a novel method based on sparse coding for image summarization, by employing both image content and tags.

(2) The similarity-inducing regularizer encourages the summary images representative in both visual and textual spaces. To the best of our knowledge, no such regularization has been proposed.

(3) Extensive experiments demonstrate our model outperforms the state-of-the-art approaches.

Related Work

The clustering method is always implemented in image summarization. In these approaches, the center images of clusters are selected as the summary. Jaffe et al. (2006) raise a problem of selecting a summary set of photos from a large collection of geo-referenced photographs available online. With the photograph locations, the authors apply the Hungarian clustering algorithm to organize images into a hierarchical structure. Simon, Snavely, and Seitz (2007) propose the problem of scene summarization, aiming to find a collection presenting the most informative aspects of the scene with minimal redundancy. To extract the representative view of a scene, the system applies clustering techniques to partition the image set into groups and then compute textual tag information that best represents each view. Similarly, (Tan et al. 2012) first automatically clusters images according to their correlation, taking similarity between various visual features into account, and selects representative images for each cluster.

Other work investigates the problem of image summarization in some unique viewpoints. In (Shi et al. 2009), the salient region, meaning foreground objects with rare occurrence and geometric constraints, is considered as an reliable description for the image summarization. Yang, Shen, and Fan (2011) treat image summarization as the problem of dictionary learning for sparse coding. The sparsity model reconstructs all images in the original collection by using a subset of most representative images. Unlike the previous work, (Sinha 2011) aims to summarize personal photos hosted on photo archives and social sharing platforms. In this work, an effective subset summary is assumed to satisfy three desirable properties, Quality, Diversity and Coverage.

Proposed Model

Preliminary

We define $\mathbf{X} = [\vec{x}_1|\cdots|\vec{x}_n] \in \mathbb{R}^{F\times n}$ as the original image set, where $\vec{x}_i$ is the image content vector, usually represented by Bag-of-Visual-Words model, and F is the feature size. Assume we have a dictionary of V possible tags $\mathcal{V} = \{w_1, \cdots, w_V\}$, then each image has a tag set, called "**image description**", denoted by $\mathcal{T}_i \subset \mathcal{V}$. Our goal is to find a subset of $\mathbf{X}$ as the summary to represent the original image set.

Joint Framework

We propose a joint framework for image summarization. The schematic illustration is shown in Figure 2. Assume $\vec{\alpha}$ and $\vec{\beta} \in \mathbb{R}^n$ are the **indicator vectors**, measuring the "representativeness" of each image from the visual and textual viewpoints respectively. Our general strategy is to simultaneously estimate $\vec{\alpha}$ and $\vec{\beta}$ via an optimization method, and select the images whose corresponding elements in $\vec{\alpha}$ and $\vec{\beta}$ are non-zero, as the summary set.

Our basic assumption is: a representative image should contain some typical tags. As a result, $\vec{\alpha}$ and $\vec{\beta}$ interact with each other. We establish the initial model as:

$$\min_{\vec{\alpha},\vec{\beta}} \; v \cdot L(\vec{\alpha};\mathbf{X}) + (1-v)\cdot L(\vec{\beta};\mathbf{M}) + \delta \cdot \Omega(\vec{\alpha},\vec{\beta}). \quad (1)$$

Here we introduce a similarity-inducing regularizer $\Omega(\vec{\alpha}, \vec{\beta})$, encouraging the summary images representative in both visual and textual spaces. $L(\vec{\alpha}; \mathbf{X})$ and $L(\vec{\beta}; \mathbf{M})$ are loss functions on image content and descriptions, where $\mathbf{M}$ is an image-tag matrix. To obtain $\mathbf{M}$, we first represent each image with their existing tags via a sparse vector $\vec{t}_i \in \mathbb{R}^V$. Then we enrich more tags to each image to gain $\mathbf{M}$ by *low-rank matrix completion*. To define $L(\cdot)$, we leverage the *reconstruction error*, a widely-used form in both document (He et al. 2012) and image summarization (Yang et al. 2012). $v \in [0, 1]$ is the linear combination coefficient.

Since the summary is a small subset of the original image set, most elements in $\vec{\alpha}$ and $\vec{\beta}$ should be zero. We impose ℓ_1 penalty to yield sparse solution. The objective function becomes:

$$\begin{aligned} \min_{\vec{\alpha},\vec{\beta}} \quad & v \cdot L(\vec{\alpha}; \mathbf{X}) + (1 - v) \cdot L(\vec{\beta}; \mathbf{M}) \\ & + \delta \cdot \Omega(\vec{\alpha}, \vec{\beta}) + \lambda \cdot (\|\vec{\alpha}\|_1 + \|\vec{\beta}\|_1). \end{aligned} \tag{2}$$

$\lambda \geq 0$ is a tuning parameter to adjust the sparseness. A large λ induces a sparse solution which leads a concise summary set. On the contrary, a small λ allows more non-zero elements, which enhance the reconstruction ability of the summarization.

In the following sections, we will demonstrate in detail: (1) the definition of loss function $L(\cdot)$ on image content and descriptions respectively; (2) the definition of the similarity-inducing regularizer $\Omega(\cdot)$ to pick out summary images representative in both visual and textual spaces; (3) an effective algorithm to solve the optimization problem.

Loss Function on Image Content

Image Set Representation Before discussing our model, we first measure the original image set as the same scale of a single image's content feature. Motivated by multi-document summarization which tends to integrate a corpus as a large document, we use a content vector $\tilde{x} \in \mathbb{R}^F$ to represent the image set $\mathbf{X}$. In Bag-of-Visual-Words representation, an image is regarded as a "visual document" and the j^{th} element of $\vec{x}_i$ denotes the weight of the j^{th} "visual word". Thus the average weight of the j^{th} visual word for the whole set is $\frac{1}{n}\sum_i x_i^j$. The **image content representation vector** is defined as follows:

$$\tilde{x} = \frac{1}{n} \sum_i \vec{x}_i. \tag{3}$$

The original image set $\tilde{x}$ can be approximated by the summary given a reconstruction function f: $\tilde{x} \approx f(\vec{\alpha}; \mathbf{X})$. Accordingly, we define the loss function of image content as the **reconstruction error**:

$$L(\vec{\alpha}; \mathbf{X}) = \|\tilde{x} - f(\vec{\alpha}; \mathbf{X})\|_2^2. \tag{4}$$

Here we define the estimator $f(\cdot)$ as a linear function $f(\vec{\alpha}; \mathbf{X}) = \sum_{i=1}^{n} \alpha_i \cdot \vec{x}_i$, where only images with non-zero α_i are involved to reconstruct the original set.

Loss Function on Image Descriptions

When image tags are applied, we expect to find which images contain representative descriptions.

Description Vectorization Similar to the image content, each description $\mathcal{T}_i$ is represented by a "textual feature" $\vec{t}_i \in \mathbb{R}^V$, whose elements are defined as

$$t_i^j = \begin{cases} \frac{1}{|\mathcal{T}_i|} & \text{if } w_j \in \mathcal{T}_i \\ 0 & \text{if } w_j \notin \mathcal{T}_i \end{cases}. \tag{5}$$

However, such textual features fail to capture the relationship occurring among different tags. For example, image A has tag "*sunset*" and image B has tag "*sundown*", so the content of two images should be very similar. However, such synonym relation cannot be recognized by the above vector representation strategy because the dimensions are assumed to be independent.

To address this problem, the intuitive idea is to enrich the tags for each image. We suppose two tags are relevant if they co-occur frequently in the descriptions. "*snow*" and "*ice*", for example, tend to be annotated simultaneously, hence we add tag "*ice*" to the images only tagged with "*snow*", i.e. assign a positive value to the corresponding dimension. We expect to recommend relevant tags to each image, and assign the weights according to the strength of correlation.

Based on the above analysis, we introduce the technology of matrix completion to enrich image tags. Cai, Candès, and Shen (2010) formalize the problem of recovering a matrix from a sampling of its entries. Let $\mathbf{M} \in \mathbb{R}^{V \times n}$ be the tag matrix after recovery. Taking inspiration from the idea of matrix completion, we suppose the observed tag matrix $\mathbf{T}$ are randomly sampled from the complete tag matrix $\mathbf{M}$. According to (Cai, Candès, and Shen 2010), $\mathbf{M}$ can be obtained by:

$$\begin{aligned} \min \quad & \tau \cdot \|\mathbf{M}\|_* + \frac{1}{2}\|\mathbf{M}\|_F \\ \text{s.t.} \quad & \forall\, \mathbf{T}_{ij} \neq 0, \mathbf{M}_{ij} = \mathbf{T}_{ij}, \end{aligned} \tag{6}$$

where $\|\cdot\|_*$ is the nuclear norm of a matrix and $\|\cdot\|_F$ is the Frobenius norm to avoid overfitting. Formula (6) can be understood as the convex relaxation of a rank minimization problem.

After tag recovery, we complete the definition of the reconstruction error on image descriptions similar to Formula (4):

$$L(\vec{\beta}; \mathbf{M}) = \|\tilde{m} - \mathbf{M}\cdot\vec{\beta}\|_2^2, \tag{7}$$

where $\tilde{m} = \frac{1}{n}\sum_i \vec{m}_i$ is the representation vector of image descriptions after tag enrichment. $\vec{m}_i \in \mathbb{R}^V$ is the i^{th} column of $\mathbf{M}$.

Similarity-inducing Regularizer

We assume visual representative images should also have typical textual descriptions. Consequently, $\vec{\alpha}$ and $\vec{\beta}$ are consistent as shown in Figure 3. One can see that for every image, the connected visual and textual indicators α_i and β_i should produce similar values.

We introduce a **similarity-inducing regularizer** $\Omega(\vec{\alpha}, \vec{\beta})$ to measure the above pairwise relations. We discuss two types of $\Omega(\vec{\alpha}, \vec{\beta})$.

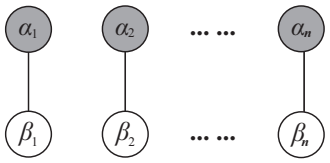

Figure 3: Pairwise relations in $\vec{\alpha}$ and $\vec{\beta}$.

Group Penalty If we regard every pair of (α_i, β_i) as a group of variables, we define $\Omega(\vec{\alpha}, \vec{\beta})$ as a group penalty:

$$\Omega_g(\vec{\alpha}, \vec{\beta}) = \sum_{i=1}^{n} \sqrt{\alpha_i^2 + \beta_i^2}. \tag{8}$$

Denote $\vec{\theta} = \begin{bmatrix} \vec{\alpha} \\ \vec{\beta} \end{bmatrix} \in \mathbb{R}^{2n}$. It is easy to see that $\Omega_g(\vec{\alpha}, \vec{\beta})$ can be written as the form of group lasso proposed in (Yuan and Lin 2006):

$$\Omega_g(\vec{\alpha}, \vec{\beta}) = \Omega_g(\vec{\theta}) = \|\vec{\theta}\|_g = \sum_{i=1}^{n} \|\vec{\vartheta}_i\|_2, \tag{9}$$

where $\vec{\vartheta}_i = \begin{bmatrix} \theta_i \\ \theta_{i+n} \end{bmatrix} = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}$. In group lasso process, the indicators from the same group, i.e. $\{\alpha_i, \beta_i\}$, prefer to have large, small, or zero weights together.

Difference-sparsity Penalty An alternative way to induce the similarity of $\vec{\alpha}$ and $\vec{\beta}$ is to penalize the difference between them:

$$\Omega_d(\vec{\alpha}, \vec{\beta}) = \sum_{i=1}^{n} |\alpha_i - \beta_i| = \|\vec{\alpha} - \vec{\beta}\|_1. \tag{10}$$

We interpret this regularizer from the sparseness point of view. Let $\vec{d} = \vec{\alpha} - \vec{\beta}$ be the difference, then the ℓ_1 penalty $\Omega_d(\vec{\alpha}, \vec{\beta}) = \|\vec{d}\|_1$ encourages $\vec{d}$ to be sparse. Therefore, most pairs of α_i and β_i share the same value due to the sparseness of their difference vector $\vec{d}$.

Algorithm

We employ the algorithm of *smoothing proximal gradient descent (SPG)* proposed by Chen et al. (2012) to solve Problem (2). We first transform it to the uniform variable expression:

$$\min_{\vec{\theta}=\begin{bmatrix} \vec{\alpha} \\ \vec{\beta} \end{bmatrix} \in \mathbb{R}^{2n}} \frac{1}{2}\|\vec{y} - \mathbf{D} \cdot \vec{\theta}\|_2^2 + \lambda \cdot \|\vec{\theta}\|_1 + \delta \cdot \Omega(\vec{\theta}). \tag{11}$$

Here $\vec{y} = 2 \cdot \begin{bmatrix} v \cdot \tilde{x} \\ (1-v) \cdot \tilde{m} \end{bmatrix} \in \mathbb{R}^{F+V}$ and $\mathbf{D} = 2 \cdot \begin{bmatrix} v\mathbf{X} & \mathbf{0} \\ \mathbf{0} & (1-v)\mathbf{M} \end{bmatrix} \in \mathbb{R}^{(F+V)\times 2n}$. Then we rewrite two types of similarity-inducing regularizer $\Omega(\vec{\theta})$.

According to (Chen et al. 2012), the group penalty can be

$$\Omega_g(\vec{\theta}) = \sum_{i=1}^{n} \sqrt{\theta_i^2 + \theta_{i+n}^2} = \max_{\vec{a} \in \mathcal{Q}} \vec{a}^T \cdot \vec{\theta}, \tag{12}$$

where $\vec{a}$ is an auxiliary vector, and the domain $\mathcal{Q} = \{\vec{a} \in \mathbb{R}^{2n} | \forall i \in \{1, \cdots, n\}, \sqrt{a_i^2 + a_{i+n}^2}\}$ is the Cartesian product of unit balls in Euclidean space.

Also, we reformulate the difference-sparsity penalty as

$$\Omega_d(\vec{\theta}) = \sum_{i=1}^{n} |\theta_i - \theta_{i+n}| = \|\mathbf{C} \cdot \vec{\theta}\|_1, \tag{13}$$

where $\mathbf{C} = [\mathbf{I}_{n\times n} | -\mathbf{I}_{n\times n}] \in \mathbb{R}^{n\times 2n}$. The inputs of SPG algorithm are the right terms of Formula (12) and (13).

The entire process of our model is shown in Algorithm 1.

Algorithm 1 JOINT model for Image Summarization

Input: Images with content features $\mathbf{X} = [\vec{x}_1 | \cdots | \vec{x}_n] \in \mathbb{R}^{F\times n}$ and the user-annotated tag matrix $\mathbf{T} = [\vec{t}_1 | \cdots | \vec{t}_n] \in \mathbb{R}^{V\times n}$.
Output: Summary set.

```
Compute the enrichment tag matrix M in Formula 6.
Calculate the representation vector x̃ and m̃.
Obtain optimal indicator vectors α⃗ and β⃗ by solving the joint summarization Problem (2), using SPG algorithm.
for i = 1 to n
    if α_i > 0 and β_i > 0, then add image i to summary set.
end for
```

Experiment

In this section, we present the experimental results of our model compared with other baseline approaches.

Experimental Setup

Data Set We leverage Yahoo! Webscope dataset [1] as the benchmark dataset in our experiments. Yahoo! Webscope dataset contains 10 categories. As each category has 200,000 images, the number of images in the entire set is 2,000,000. The images in the dataset are represented as the features in a bag-of-words format using 400 codewords. According to the URLs of each image, we grab the tags on the webpage to form the image description.

Baseline Methods In order to demonstrate the effectiveness of our model, we have selected four baseline algorithms for comparison.

- *ARW* (Zhu et al. 2007): ARW is a ranking algorithm based on random walks in an absorbing Markov chain, with an emphasis on diversity. The ranked items are turned into absorbing states, which prevent redundant items from receiving a high rank.

[1] Yahoo! Webscope dataset ydata-flickr-ten-tag-images-v1_0 http://labs.yahoo.com/Academic_Relations

	ID	Original description	Recommended tags
High Accuracy	1	China Beijing Tiananmen	Peking city Shanghai China2007 Asia
	2	vacation Europe France mom	Paris trip parents travel06 Paris11
	3	squash lasagne cooking dinner	kitchen Italian eating chef friends
Low Accuracy	4	celebration groom bride love family marriage	flowers church chocolate cake party
	5	Eastlondon BromleybyBow geotagged Threemills	warehouse film wall Thames wharf

Table 1: Tag Recommendation. The font size of each recommended tag indicates the weight to the description.

		nature	food	sky	travel	2012	beach	London	music	people	wedding	*Average*
Our Model	**JOINT-dif**	**0.62**	0.51	**0.57**	**0.46**	0.40	**0.73**	**0.50**	**0.56**	**0.50**	**0.38**	***0.52***
	JOINT-group	0.58	**0.53**	0.43	0.32	**0.47**	0.42	0.33	0.38	0.25	0.33	*0.40*
Baselines	ARW	0.42	0.38	0.18	0.25	0.44	0.54	0.37	0.50	0.44	0.29	*0.38*
	k-medoids	0.42	0.32	0.20	0.25	0.31	0.33	0.21	0.31	**0.50**	0.17	*0.30*
	(Yang et al. 2012)	0.42	0.34	0.28	0.29	0.36	0.29	0.29	0.37	0.44	0.29	*0.34*
	SDS	0.29	0.28	0.39	0.29	0.30	0.33	0.25	0.44	0.44	**0.38**	*0.34*

Table 2: Subjective Evaluation. Best precisions are shown in bold.

- *k-medoids* (Hadi, Essannouni, and Thami 2006): k-medoids algorithm (Kaufman and Rousseeuw 1987) is a typical partitioning clustering algorithm similar to k-means, but chooses datapoints as centers. Hadi, Essannouni, and Thami (2006) propose a summarization algorithm based on the k-medoids clustering to find the best representative frame for video shots. In image summarization, the algorithm chooses the center of each cluster as summary images.
- (Yang et al. 2012): Yang et al. (2012) formulate the summarization problem as an issue of dictionary learning for sparse representation. The algorithm selects bases which can be sparsely combined to represent the original image and achieve a minimum global MSE (Mean Square Error). To avoid the local optimum, a simulated annealing algorithm is adopted to minimize the objective function.
- *SDS* (Krause and Cevher 2010): SDS represents a series of greedy algorithm which iteratively select the current best basis. The greedy approximation algorithms are guaranteed to achieve a theoretical lower bound, as the objective function satisfies submodularity.

Preprocessing

Before image summarization, we enrich the image tags by matrix completion. Table 1 shows example descriptions with their enriched tags where "recommended tags" are top-5 weighted tags for each image. The table shows 2 kinds of results, reliable recommendations and irrelevant recommendations.

From the "high accuracy" group we find matrix completion is able to supplement tags highly correlated to the original description, e.g. "*Peking*" for the 1^{st} description, "*parents*" for the 2^{nd} description, and "*cooking*" and "*chef*" for the 3^{rd} description. Matrix completion can recognize topic-related tags as well, e.g. "*trip*" for the 2^{nd} description, and "*eating*" for the 3^{rd} description. However, when the original tags are infrequent terms, such as "*BromleybyBow*" and "*Threemills*" in the 5^{th} description, matrix completion fails to append trustworthy tags.

Subjective Evaluation

In this experiment, we evaluate the performance of each method. We first generate original image collections from Yahoo! Webscope dataset. Then unbiased assessors are asked to pick their own summaries as ground truth. We finally compare the output of each method and the ground truth, and report the precisions.

Establish ground truth We randomly pick image sets from every category of Yahoo! Webscope dataset as the original image collections to be summarized. Since every category of the dataset contains 200,000 images, even within the same category, images express a plenty of themes. It is obvious that summarization algorithms make sense only if the original image collection includes a small number of topics. For instance, there is no need to summarize 1,000 travel photos with 800 tourist sites, because it is impractical to find a small amount of subset to cover all sites. In order to control the topic number of the image collection, we require the selected images contain specific key tags. For every category, we pick the image collection as follows:

(1) Randomly select 10 key tags according to the occurrence frequency in the category;

(2) Select images whose descriptions contain at least one of 10 key tags to form the collection;

(3) If more than 100 images are picked by (2), uniformly pick 100 images among them at random to form the image collection. If less than 50 images are chosen, go to (1) to repick the collection.

The process is repeated until we pick 5 image collections for each category. On average, we harvest 82 images for each collection.

To establish ground truth we ask 8 assessors to select summaries on these collections based on their personal preferences. The assessors are allowed to pick 10 images from each collection. The images chosen at least 3 times are selected as the ground truth.

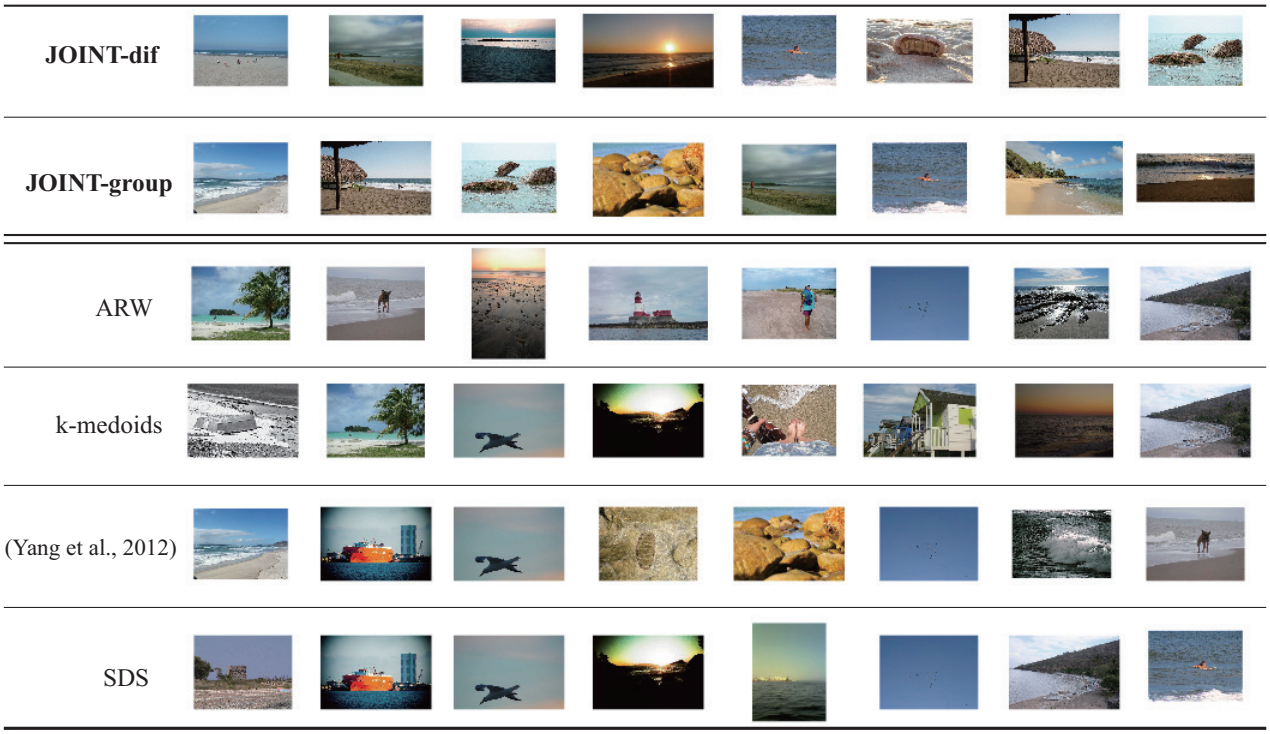

Figure 4: Case study. Summaries of all methods on a beach image collection.

Result Each method outputs 8 images as the summary. In Table 2, we report the average matching rate between the summary set of each method and the ground truth for every category, measured by *precision*.

Our two models (JOINT-dif and JOINT-group) achieve best results in all categories and top-2 average precisions, which indicates our approach dominates the baseline methods. Especially on the categories like "*nature*", "*food*", "*sky*" and "*travel*", JOINT-dif and JOINT-group achieve two highest accuracies. We find that images of these four categories tend to express only a small amount of topics. Therefore, images with topic tags are very likely to be selected as a summary. On the contrary, our methods do not show great advantages on the categories like "*2012*", "*people*" and "*wedding*". Similar to the above analysis, we guess the themes in these categories are vague. For example, the images in "*2012*" may be tourist photos, dinner pictures or images of some incidents occurred in 2012. We also note that JOINT-dif, attaining best results in 8 categories, outperforms JOINT-group. This suggests that the difference-sparsity penalty is more specifically suited than the group penalty to induce similarity between two summary indicators.

Among four baselines, ARW outperforms the other three on most categories, indicating diversity is an important factor in image summarization.

Case Study

In order to illustrate the characteristics of six methods, we display the summaries of all methods on a beach image collection in Figure 4. From the first two rows, we can see that the summaries of our approaches, covering pictures depicting beaches in different times or views, show diversity well. In contrast, the summaries generated by the baselines tend to comprise redundant images, such as the 4^{th} and 5^{th} picture of (Yang et al. 2012), or the 3^{rd} and 6^{th} picture of (Yang et al. 2012) and SDS. Furthermore, the baseline methods are inclined to simultaneously choose the images that our methods do not choose, like the 3^{rd} image of k-medoids, (Yang et al. 2012) and SDS. It might be because that these images, whose visual features are close to clustering centers in Euclidean space, lack key tags in their user-annotated description.

Conclusion

In this paper, we propose an effective model based on sparse coding to select the images with both representative visual content and textual descriptions from a large-scale image collection. To leverage the image tags, we introduce two types of similarity-inducing regularizer: group penalty and difference-sparsity penalty. L_1 penalty makes a sparse solution since the summary set is a small subset of the original collection. The experiment results demonstrate our model outperforms the state-of-the-art approaches and classical summarization methods. Moreover, JOINT-dif achieves the higher average precision than JOINT-group, suggesting the difference-sparsity penalty is a better choice to capture the similarity between the visual and textual indicator vectors. From the experiment of the case study, one can see that by considering the human annotated tags, we significantly im-

prove the diversity, quality and relevancy of the summary.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 61170091).

References

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Chen, X.; Lin, Q.; Kim, S.; Carbonell, J. G.; and Xing, E. P. 2012. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics* 6(2):719–752.

Hadi, Y.; Essannouni, F.; and Thami, R. O. H. 2006. Video summarization by k-medoid clustering. In *Proceedings of the 2006 ACM symposium on Applied computing*, 1400–1401. ACM.

Hafner, J.; Sawhney, H. S.; Equitz, W.; Flickner, M.; and Niblack, W. 1995. Efficient color histogram indexing for quadratic form distance functions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 17(7):729–736.

He, Z.; Chen, C.; Bu, J.; Wang, C.; Zhang, L.; Cai, D.; and He, X. 2012. Document summarization based on data reconstruction. In *AAAI*.

Jaffe, A.; Naaman, M.; Tassa, T.; and Davis, M. 2006. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, 89–98. ACM.

Kaufman, L., and Rousseeuw, P. 1987. Clustering by means of medoids.

Krause, A., and Cevher, V. 2010. Submodular dictionary selection for sparse representation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 567–574.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, 1150–1157. Ieee.

Shi, L.; Wang, J.; Xu, L.; Lu, H.; and Xu, C. 2009. Context saliency based image summarization. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 270–273. IEEE.

Simon, I.; Snavely, N.; and Seitz, S. M. 2007. Scene summarization for online image collections. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 1–8. IEEE.

Sinha, P. 2011. Summarization of archived and shared personal photo collections. In *Proceedings of the 20th international conference companion on World wide web*, 421–426. ACM.

Tan, L.; Song, Y.; Liu, S.; and Xie, L. 2012. Imagehive: Interactive content-aware image summarization. *Computer Graphics and Applications, IEEE* 32(1):46–55.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Wang, J.; Jia, L.; and Hua, X.-S. 2011. Interactive browsing via diversified visual summarization for image search results. *Multimedia systems* 17(5):379–391.

Yang, J.; Jiang, Y.-G.; Hauptmann, A. G.; and Ngo, C.-W. 2007. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 197–206. ACM.

Yang, C.; Shen, J.; Peng, J.; and Fan, J. 2012. Image collection summarization via dictionary learning for sparse representation. *Pattern Recognition*.

Yang, C.; Shen, J.; and Fan, J. 2011. Effective summarization of large-scale web images. In *Proceedings of the 19th ACM international conference on Multimedia*, 1145–1148. ACM.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Zhu, X.; Goldberg, A. B.; Van Gael, J.; and Andrzejewski, D. 2007. Improving diversity in ranking using absorbing random walks. In *HLT-NAACL*, 97–104.