Automatic Image Tagging via Category Label and Web Data

Shenghua Gao,   Zhengxiang Wang,   Liang-Tien Chia,   Ivor Wai-Hung Tsang
School of Computer Engineering, Nanyang Technological University,Singapore
{gaos0004,wang0460, asltchia, IvorTsang}@ntu.edu.sg

ABSTRACT

Image tagging is an important technique for the image content understanding and text based image processing. Given a selection of images, how to tag these images efficiently and effectively is an interesting problem. In this paper, a novel semi-auto image tagging technique is proposed: By assigning each image a category label first, our method can automatically recommend those promising tags to each image by utilizing existing vast web data. The main contributions of our paper can be highlighted as follows: (i) By assigning each image a category label, our method can automatically recommend other tags to the image, thus reducing the human annotation efforts. Meanwhile, our method guarantee tags' diversity due to abundant web data. (ii) We use sparse coding to automatically select those semantically related images for tag propagation. (iii) Local & global ranking agglomeration will make our method robust to noisy tags. We use Event dataset as the images to be tagged, and crawled Flickr images with their associated tags according to the category label in Event dataset as the auxiliary web data. Experimental results show that our method achieves promising performance for image tagging, which proves the effectiveness of our method.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Selection process

General Terms

Algorithm, Experiment, Performance

Keywords

Category Label, Web Data, Automatic Image Tagging.

1. INTRODUCTION

Tagging helps users and systems to better understand the semantic content of media. It is the basis of many text based image processing techniques, such as text based image retrieval [7], text aided image classification [13] *etc.* However, manual image tagging requires time and effort, and for such a tedious process, it is difficult for image owners/users to annotate all relevant tags for each image.

Given a scenario: a photographer takes lots of photographs while he attends some activities(such as a school sports day), or an event(wedding, birthday celebration,*etc.*). When he is ready to organize his collection of images, he will need comments/tags to be added to these images, and such tags will greatly facilitate image retrieval in the future. If he is sharing those images on online photo communities/forums, tags can help visitors to better understand and navigate his image collection. The issue here is: Is there some better image tagging solution than manually annotating each image? Will the solution provide an efficient and effective method for annotating images?

On the one hand, we know that the images to be annotated belong to only limited number of categories. Here 'category' means those higher semantic level words which are commonly used for describing the images with some similar properties, such as the categories defined in NUS-WIDE dataset[2]. For instance, if the images are taken in some event/activities, these images may belong to 'badminton', 'swimming', 'wedding', *etc.* If these images are taken in a trip, these images may belong to 'tiger', 'building', 'mountain', *etc.* It is hard to assign all the tags to each image, but it is easy to assign each image a category label. For example, the images in the same category are usually sequentially located in image folder, so one can easily add a category label to these images. On the other hand, billions of web images with their associated tags are available currently. These web data provide plentiful candidate tags for users, which can be used to solve our image tagging problem.

Motivated by these two reasons, a new image tagging technique is proposed. Our method requires users to annotate each image with ONE category label first. By using these category labels, classifiers can be constructed to get rid of those content irrelevant web images, the remaining of relatively clean images form the codebook images set used for tag propagation. Then sparse coding technique is used to select a small subset of relevant images from the codebook images, meanwhile the corresponding tags of these relevant images form a tag pool. By utilizing local & global ranking agglomeration technique, those most promising tags are automatically selected and recommended to the image to be tagged. In this way, the human labor needed in the image tagging process is greatly minimized. Experimental results

MM'10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

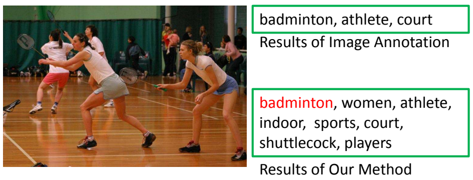
badminton, athlete, court
Results of Image Annotation
badminton, women, athlete, indoor, sports, court, shuttlecock, players
Results of Our Method

Figure 1: Illustration of the difference between our method and image annotation. Different tags can be used to describe the image content in different views. For example, 'athlete', 'player', 'woman', 'indoor' all contribute to the image content description('badminton' is the category label user assigned).

show that the recommended tags can enhance the image content description.

The rest of this paper is organized as follows: In Section 2, we review two systems that are most closely related to our work: image annotation system and tag recommendation system. By analyzing the differences between our work and these two systems, we highlight the main contributions of our work. In Section 3, the main modules of our method are given in details. The evaluation of our method and experimental setup are given in Section 4. In the end, we summarize our work in Section 5.

2. RELATED WORK AND OUR CONTRIBUTIONS

In image annotation systems [8, 12], the lexicon of tags is pre-defined. Classifiers are trained according to the tags in lexicon. When a new image comes in, the system decides which tags should be assigned to it by using the pre-trained classifiers. There are some obvious drawbacks in image annotation system. First of all, the tags assigned to the new image cannot exceed the pre-defined lexicon, which constrains the tags' diversity(Fig. 1). However, in our method, we use vast number of web images and their tags for image tagging, which guarantees the diversity of tags. Secondly, in image annotation problem, training data are required to be manually annotated which is a tedious and cost expensive work. Meanwhile, the number of classifiers equals to the the number of tags, which increases the computational cost. Our method only needs user to annotate one category label, and train a classifier for each category label, whose number is much less than that of tags, so the computational cost is also reduced.

Another system related to our work is tag recommendation system [9, 10]. In a image tag recommendation system, several tags are randomly initialized beforehand for each image. Based on the given tag and visual feature, some tags are recommended to this image. For tag recommendation system, each image is tagged independently, so it can easily be affected by noisy web images. However, in our method, we use the images with same category label to train a classifier to filter out those noisy web images. So our method uses the information among the images to be tagged to enhance its robustness to noise, which improves the system's performance. This is the biggest difference of our system and tag recommendation system. Moreover, labeling one category label is much easier than listing several tags.

The contributions of this paper can be summarized as follows: (i) Our method uses category label to aid image tagging, which can filter out those noisy content irrelevant images and reduce the human efforts. Moreover, Our method boosts the diversity of the tags. (ii) Sparse coding technique is used to select those semantic relevant images for tag propagation, and it is also robuster to noise than ranking based kNN method [1]. (iii) Local & global ranking agglomeration technique is used to reduce the tag noisy level for more accurate image tagging.

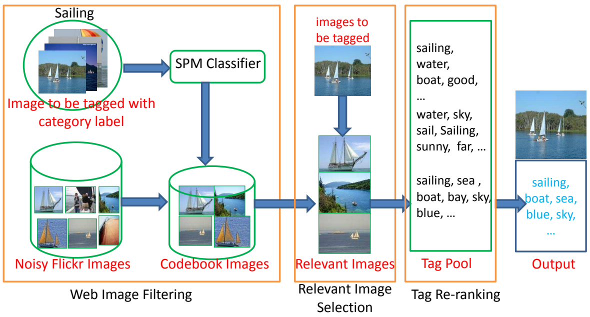
Sailing
Image to be tagged with category label
SPM Classifier
Noisy Flickr Images
Codebook Images
Web Image Filtering
images to be tagged
Relevant Images
Relevant Image Selection
sailing, water, boat, good, ...
water, sky, sail, Sailing, sunny, far, ...
sailing, sea , boat, bay, sky, blue, ...
Tag Pool
Tag Re-ranking
sailing, boat, sea, blue, sky, ...
Output

Figure 2: Flowchart of our automatic image annotation

3. CATEGORY LABEL AND WEB DATA AIDED IMAGE TAGGING

As shown in Fig. 2, our method contains three modules: (i) Noisy Images Filtering: Spatial Pyramid Matching(SPM) [4] classifiers are constructed to filter out those content irrelevant web images. (ii) Relevant Image Selection: Sparse coding is used to select those semantically related images for tag propagation. (iii) Tag Re-ranking: We adopt local & global ranking agglomeration method to re-rank all the tags corresponding to the relevant images, and use those most promising tags for image tagging.

3.1 Noisy Images Filtering

Web images downloaded according to certain keyword query word are very noisy: lots of returned images are content irrelevant to the query. Directly using these noisy web images may worsen the performance of tag propagation, so it is necessary to get rid of irrelevant web images. Different from image annotation and image recommendation systems, images are annotated with a category label in our method. Thus good classifiers can be trained by taking advantage of these clean data. With the help of these classifiers, we can re-rank the noisy web images.

Recently lots of research has been carried out on image classification, in which SPM[4] has shown its good performance in preserving spatial information in Bag-of-Word(BoW) based image representation. In our method, we adopt SPM to represent images. Specifically, for a given category, we use the images with the same label and images with different label as positive and negative instances respectively to train a one-vs.-all SVM classifier. Histogram intersection kernel is used. Then the web images are re-ranked according to their decision value output of SVM classifier. The images with higher decision values are more likely to be positive images for the given category. We only select these images to form the codebook image for next module. In this module, SPM classifier not only decreases the image noisy level, but also reduces the number of web images, which decreases the computational cost for sparse coding in Section 3.2.

3.2 Relevant Image Selection

The performance of automatic image tagging relies heavily on the selection of relevant web images used in tag propagation process. Recent research has shown the effectiveness of sparse coding(Sc) in solving computer vision problems, such as image annotation[12], label propagation[3, 6]. Following these works, we also adopt sparse coding for relevant web image selection.

Suppose the image to be tagged is represented by y, and the codebook images, which are the top K images ranked by their decision value, are: $X = [x_1, x_2, \ldots, x_K]$. Sparse coding automatically selects as few semantically related instances as possible from codebook images X to linearly reconstruct y, and the coefficients corresponding to X are $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K]$. Under the sparse constraint, only a small fraction entries in α are non-zero. We formulate the sparse coding problem as follows:

$$\underset{\alpha}{min} \;\; \|y - X\alpha\|_F^2 + \lambda\|\alpha\|_1 \tag{1}$$

In Eq. (1), the first term is the reconstruction error, and the second term is the sparse constraint. λ is the trade-off parameter between the reconstruction error and sparsity. Empirically larger λ corresponds to sparser solution. In our experiment, we fix λ to 0.2. Then the relevant image set corresponding to y which is the image to be tagged, can be defined as follows by using ℓ^1 Nearest Neighbor [11]:

$$X_r = \{x_m | \alpha_m > 0, m \leq K\} \tag{2}$$

Compared with kNN method, sparse coding can exploit the semantic relationship between different images [12]. For example, in sparse coding, a rockclimbing image may be linearly reconstructed by images containing rock and images containing person, though the distances between the images of rockclimbing, rock, person may be large. In other words, sparse coding aims at using different parts from the different images to linearly construct the objective image. In this sense, sparse coding can make use of more information than kNN method. Moreover, kNN is a ranking based method [1] and sensitive to the noise. But in sparse coding, the noise can be reduced to some extent, because we do not consider the ranking in the relevant image set.

3.3 Tag Re-ranking

The raw tags of codebook images are very noisy, which will greatly harm the final tagging performance. So tag denoising is necessary. To achieve this, we first use WordNet[1] to get rid of those misspelt words. Then we filter out the irrelevant tags such as country names, person names and stopwords like ‘is’, ‘an’, ‘there’ *etc.* Tags whose frequency is less than 2 are also removed. The remaining tags in codebook images form the tag pool L.

Suppose there are M images in relevant image set: $X_r = [x_{m_1}, x_{m_2}, \ldots, x_{m_M}]$. The tags corresponding to these images are $T = \{T_{m_1}, T_{m_2}, \ldots, T_{m_M}\}$, in which T_{m*} is the tag set of image x_{m_*}. Lots of tags are duplicate in T. Suppose there are N non-duplicate tags in all in T: $T' = \{t_1, t_2, \ldots, t_N\}$. For these N tags, local & global ranking agglomerating method is used to determine the tag assignment priority.

For all the tags in T', we can assign them a global ranking score according to their global frequency in L: $R_G = [r_1^g, r_2^g, \ldots, r_N^g]$. We can also assign them a local ranking score according to the frequency in T: $R_L = [r_1^l, r_2^l, \ldots, r_N^l]$ ($r_*^l, r_*^g \in \{1, 2, \ldots, N\}$. Lower ranking scores correspond to more frequently appearing tags). It is reasonable that those globally frequently appearing tags are more prone to appear in image y, and those less frequently tags are likely to be noise. It is also easy to understand that those local frequently tags are more relevant to the images to be annotated. To balance these two important factors, we use the following equation to agglomerate the two rankings.

$$s_* = log(r_*^g) + log(r_*^l); \tag{3}$$

In which s_* is the integral score for tag t_*. Eq. (3) implies both global and local ranking scores should be small for the final recommenced tags. We re-rank all the tags based on s_* in ascend order, and those tags in the top positions will be tagged to the image accordingly.

Compared with global ranking based tag assignment, the tags assigned to the images of our method vary a lot in terms of tags' diversity. Besides this, our method achieves higher accuracy than local ranking based tag assignment by considering the global frequency of the tags.

4. EXPERIMENTS

4.1 Data Preparation

To evaluate our method, we use the Event dataset [5] as the data to be tagged. There are 8 categories in Event dataset: *badminton, bocce, croquet, rockclimbing, polo, rowing, sailing* and *snowboarding.* The image number ranges from 137 to 250 for each category and there are 1792 images in all.

We crawled 4000 images[2] from Flickr[3] by using the category label as the query. We then eliminate those images whose aspect ratio is larger than 2 or less than 0.5. After that, the image number for each category is around 3000-3500. We use the first 1500 images of the output of SVM classifier as the the codebook images.

4.2 Experimental Setup

SIFT has been widely used for image representation. In our method, we adopt SIFT feature in SPM. Specifically, we resize the maximum side(length/width) of each image to 400 pixels, and aspect ratio is kept. We fix the step size and patch size to 8 and 16 respectively. We quantize the SIFT into 400 cluster centers in k-means. In SPM, we set the layer level to 3.

Texture and color are two important aspects for predicting image content. Following the work in ref [14], we use gist and block-wise color histogram in HSV space for image representation in sparse coding process. For gist feature, we use 3 scales, and the orientation number in each scale is 8. And each image is divided into 4 blocks. For color histogram, we quantize H into 16 bins, and quantize S and

[1] http://wordnet.princeton.edu/

[2] Some images are not available due to copyright protection or other reasons. The real number ranges from 3500-4000 for each category.

[3] http://www.flickr.com/

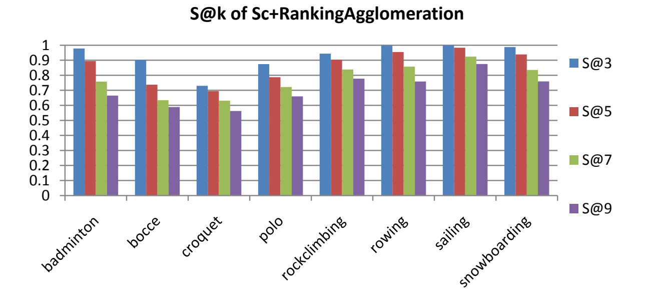

Figure 3: S@k of our method for different categories at different k tags .

V into 4 bins. We use 4*4 block-wise color histograms. We normalize the length of the color histogram and the gist to 1 respectively by using $\ell 2$ norm. Then we concatenate gist and HSV histogram to a long vector to represent the image.

4.3 Evaluation

Following the method in [9], we use the *success at rank k* ($S@k$) to evaluate the performance of automatic image tagging[4]. Specifically, if m tags are correct in top k recommended tags:

$$S@k = \frac{m}{k} \times 100\% \qquad (4)$$

We list the performance of our method in Fig. 3. It can be seen that for most categories, the top 3 recommended tags are all correct. Even for the top 9 recommended tags, the S@k is still above 50%. We notice that for 'bocce', 'croquet', 'polo', the performance is relatively low, which may result from the fact that the images in codebook images are still noisy for these three categories. Taking 'polo' for example, there are lots of images related to polo T-shirt, and only 2/3 of all the 1500 codebook images really relates to 'polo' (sports). This also proves the importance of web image pre-filtering.

We also compare our method with other two methods: Sparse Coding with Local Ranking(Sc+LocalRanking) and kNN based method[5][1]. Fig. 4 shows that our sparse coding and local & global ranking agglomeration based method achieves the best performance. We can also see that our sparse coding based methods outperforms kNN based method, which proves the effectiveness of sparse coding. Some results are listed in Fig. 5.

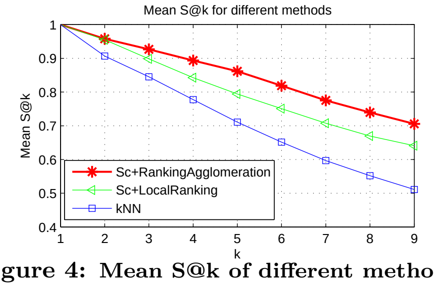

Figure 4: Mean S@k of different methods.

5. CONCLUSION

In this paper, a novel web image based automatic image tagging technique is proposed. Our method only requires users to assign each image a category beforehand, then the other tags can be automatically tagged to this image by using sparse coding technique and local & global ranking agglomeration. In this way, the human annotation efforts can be greatly reduced. Moreover, category label can also reduce the noisy level of web images for better tag propagation. Experimental results show that our method achieves very good performance and outperforms kNN based method.

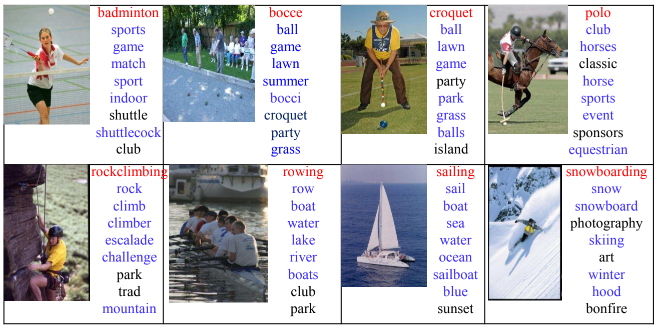

Figure 5: Some results of our method. The tags in red are labeled by users, and the tags in blue are judged as right tags.

6. REFERENCES

[1] M. Ameesh, P. Vladimir, and K. Sanjiv. A new baseline for image annotation. In *ECCV*, 2008.

[2] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[3] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[5] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[6] X. Liu, B. Cheng, S. Yan, J. Tang, T.-S. Chua, and H. Jin. Label to region by bi-layer sparsity priors. In *ACM Multimedia*, 2009.

[7] Y. Liu, D. Xu, I. W. Tsang, and J. Luo. Using large-scale web data to facilitate textual query based retrieval of consumer photos. In *ACM MM*, 2009.

[8] Z. Lu, H. H.-S. Ip, and Q. He. Context-based multi-label image annotation. In *CIVR*, 2009.

[9] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW*, 2008.

[10] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *ACM SIGIR*, 2008.

[11] C. Wang, S. Yan, and H.-J. Zhang. Large scale natural image classification by sparsity exploration. In *ICASSP*, 2009.

[12] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang. Multi-label sparse coding for automatic image annotation. In *CVPR*, 2009.

[13] G. Wang, D. Hoiem, and D. A. Forsyth. Building text features for object image classification. In *CVPR*, 2009.

[14] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and M. Dimitris. Automatic image annotation using group sparsity. In *CVPR*, 2010.

[4] In our method, two annotators are invited to judge whether the recommended tags are correct or not.

[5] In *kNN* based method, we firstly select the k nearest neighbors from codebook images for the image to be tagged, and propagate the tags of these images to this image according to each tag's frequency ranking and images' distance ranking.