Smart Batch Tagging of Photo Albums*

Dong Liu†, Meng Wang‡, Xian-Sheng Hua‡, Hong-Jiang Zhang§
† School of Computer Sci. & Tec., Harbin Institute of Technology, Harbin, 150001, P.R.China
‡ Microsoft Research Asia, Beijing, 100190, P.R.China
§ Microsoft Advanced Technology Center, Beijing, 100190, P.R.China
dongliu@hit.edu.cn,{mengwang,xshua,hjzhang}@microsoft.com

ABSTRACT

As one of the emerging Web 2.0 activities, tagging becomes a popular approach to manage personal media data, such as photo albums. However, exhaustively tagging all photos in an album is a labor-intensive and time-consuming task, and simply entering tags for the whole album will significantly degrade the tagging accuracy. In this paper, we propose a smart batch tagging scheme that aims at facilitating users in album tagging. For a given album, it selects a set of representative exemplars for manual tagging, where the number of exemplars is dependent on the content of the photos. Then the tags of the rest photos are automatically inferred. In this way, the number of tagged photos is significantly reduced and we will show that high tagging accuracy can still be maintained. Therefore, a good trade-off between manual efforts and tagging performance can be achieved. Experimental results have demonstrated the effectiveness and usefulness of the proposed approach.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Management

Keywords

Batch tagging, AP, Propagation, Flickr

1. INTRODUCTION

With the popularity of digital cameras, recent years have witnessed a rapid growth of personal photo albums. People capture photos to record their lives and share them on the web. For example, Flickr [1], the earliest and the most popular photo sharing website, hosts over 200-million personal photos.

Tagging has proved to be a successful approach to facilitate the management and the sharing of photos. By providing tags to describe the content of photos, many manipulations can be easily accomplished, such as indexing, browsing and search. The most intuitive approach to generate tags is to investigate automatic tagging (or annotation) techniques [2, 3, 4]. However, although encouraging advances have been achieved in automatic tagging technology, currently these methods can still hardly obtain satisfactory performance for real-world photos that contain highly varying content. As an example, Fig. 1(a) illustrates two photos and the tags predicted by ALIPR, a state-of-the-art image annotation system introduced in [2]. From the results we can see that many of the tags are incorrect. In fact, nowadays most photo sharing websites adopt the manual tagging approach, i.e., allowing users to manually enter tags to describe their uploaded photos. But a problem of manual tagging is its labor cost: a simple study in [5] shows that a user needs 6.8 seconds to enter a tag for an image. Therefore, exhaustively tagging all the photos in a large album will be a labor-intensive task.

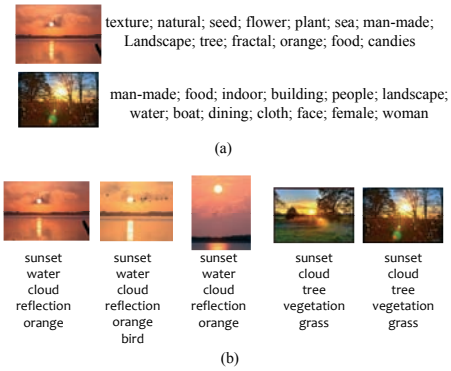

Figure 1: (a) The tags obtained by the ALIPR system [2] for two photos, and we can see that many of them are inaccurate. (b) Several photos and their tags from a personal album on Flickr. Many photos are both visually and semantically close.

To help users accomplish tagging more efficiently, in this work we introduce a smart batch tagging scheme which only needs users to manually tag several selected exemplars and

*This work was performed at Microsoft Research Asia.

MM'09, October 19–24, 2009, Beijing, China.

the tags of the rest photos are inferred automatically. This approach benefits from the fact that many photos in a personal album are usually captured continuously to record one or more events, and thus many of them are close [4], such as the examples illustrated in Fig. 1(b). Therefore, a large part of manual efforts may be redundant in the exhaustive tagging manner (i.e., tagging all photos in the album). For example, we actually only need to select and tag two photos (such as the first and the fourth photos) in Fig. 1(b) and the tags of the rest photos are not difficult to predict.

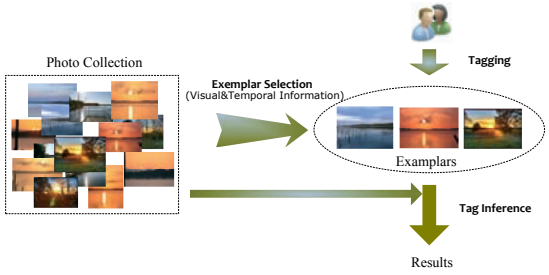

Figure 2: The schematic illustration of smart batch tagging.

Fig. 2 illustrates the schematic illustration of the proposed batch tagging approach. There are two main components in the scheme, i.e., exemplar selection and tag inference. Given an album, the photos are first grouped into a set of clusters and an exemplar is selected from each cluster. The number of clusters is dependent on the content of the photos. Users then only need to manually tag the selected exemplars, and the tags of the rest photos are obtained through an inference step. For exemplar selection, we propose a temporally consistent affinity propagation method that integrates both visual and temporal information, and for tag inference we employ a tag propagation method. Empirical study will demonstrate their superiority over many other methods. The scheme is able to significantly reduce human efforts while maintaining highly accurate tagging results, i.e., it achieves a trade-off between manual efforts and tagging performance. It can be applied in either online photo sharing or desktop photo management services.

The organization of the rest of this paper is as follows. In Section 2, we provide a short review on the related work. In Section 3 and Section 4, we introduce the adopted exemplar selection and tag inference methods, respectively. Empirical study is presented in Section 5. Finally, we conclude the paper in Section 6.

2. RELATED WORK

Extensive research efforts have been dedicated to photo tagging. Ames et al. [6] have explored the motivation of tagging on the Flickr website and they claim that most users tag photos to make them better accessible to the general public. Kennedy et al. [7] have evaluated the performance of the classifiers trained with Flickr photos and the associated tags. Liu et al. [8] have proposed a method to analyze the relevance scores of tags with respect to an image. Yan et al. [5] propose a model that is able to predict the time cost of manual image tagging. Tag recommendation is an intensively studied approach to help users tag photos more efficiently [9, 10]. By recommending a set of potentially relevant keywords in the tagging process, users can directly select the correct ones instead of entering them and it can effectively reduce labor cost. This work adopts a different approach to facilitate users in tagging. For an album, only a set of selected photos are manually tagged, and the tags of the other photos are automatically inferred. In this way, the manual efforts can be significantly reduced and we will demonstrate that fairly high tagging accuracy can still be maintained.

3. EXEMPLAR SELECTION

The exemplar selection is accomplished via a temporally consistent affinity propagation method. We first introduce Affinity Propagation (AP) [11]. It is a similarity measure-based clustering algorithm that is able to group a given set of samples into several clusters as well as select an exemplar from each cluster.

Given a set of n samples $X = \{x_1, x_2, \ldots, x_n\}$, the algorithm propagates two kinds of information among the samples: the "responsibility" $r(x_i, x_k)$ sent from x_i to x_k, which reflects how well x_k serves as the exemplar of x_i considering other potential exemplars for x_i, and the "availability" $a(x_i, x_k)$ sent from x_k to x_i, which indicates how appropriate x_i chooses x_k as its exemplar considering other potential samples that may choose x_k as their exemplar. The algorithm then works by iterating

$$r(x_i, x_k) = s(x_i, x_k) - \max_{k' \neq k}\{a(x_i, x_{k'}) + s(x_i, x_{k'})\} \quad (1)$$

$$a(x_i, x_k) = \begin{cases} min\{0, r(x_k, x_k) + \sum_{i' \notin \{i,k\}} max\{0, r(x_{i'}, x_k)\}\} & if\ i \neq k \\ \sum_{i' \neq k} max\{0, r(x_{i'}, x_k)\} & otherwise \end{cases} \quad (2)$$

where $s(x_i, x_k)$ denotes the similarity between x_i and x_k which will be formulated in Eq. 5.

After convergence, the exemplar $e(x_i)$ is decided as x_k where

$$k^* = \arg\max_k \{a(x_i, x_k) + r(x_i, x_k)\} \quad (3)$$

Denote by $e(x)$ the exemplar of x, then it can be proved that AP actually maximizes the sum of similarities between each sample and its exemplar [11], i.e.,

$$S(X, E) = \sum_{i=1}^{n} s(x_i, e(x_i)) \quad (4)$$

We choose AP in our smart batch tagging task due to its advantages in the following aspects:

- It simultaneously accomplishes the clustering of samples and the selection of exemplars. Several other methods, such as k-means and spectral clustering, merely cluster samples and the centroids of clusters may not be real samples.

- It is able to automatically determine the number of clusters, and this is particularly advantageous for the batch tagging task. For example, it is able to select more exemplars for albums that contain significantly varied photos, and contrarily select fewer exemplars for the albums that are less diverse.

Most existing works model the similarity of two images based on their visual features [12, 13]. However, time is an important information clue for personal photos [14]. Since the photos in an album are captured by the same person,

two photos that are temporally close will have high probability to record an identical scene or event. Therefore, we integrate the visual and temporal information to compute the similarities of photos. More specifically, the similarity between photos x_i and x_j is estimated as

$$s(x_i, x_j) = \alpha exp(-\frac{\|v_i - v_j\|^2}{\sigma_v^2}) + (1-\alpha)exp(-\frac{\|t_i - t_j\|^2}{\sigma_t^2}) \quad (5)$$

where v_i and t_i indicate the visual feature vector and timestamp of photo x_i respectively, and α is a weight factor between 0 and 1. AP is performed with this similarity measure, and we name this method temporally consistent AP since the selected exemplars will be not only visually representative but also cover widely and diversely in time.

4. TAG INFERENCE

After manually tagging the selected exemplars, the next step is to infer the tags of the rest photos. Denote by $\Omega = \{t_1, t_2, \ldots, t_m\}$ the set of appeared unique tags for the photo album. Denote by $\mathbf{y}(x_i)$ the tag membership vector for photo x_i, in which the k-th entry indicates the membership of tag t_k to the photo, i.e., $y_k(x_i) = 1$ if t_k is relevant to x_i and otherwise $y_k(x_i) = -1$. Specifically, $\mathbf{y}(x_i)$ corresponding to exemplars are available according to user provided tags and our task is to estimate $\mathbf{y}(x_i)$ for the other photos. The most intuitive approach is to directly assign the tags of each exemplar to the whole cluster, i.e., $\mathbf{y}(x_i) = \mathbf{y}(e(x_i))$ (we name it naive tag assignment). However, this method heavily relies on the performance of clustering and it neglects the different distances of photos to exemplars. Therefore, here we adopt a tag propagation method, which is closely related to a graph-based semi-supervised learning approach [15]. The method works by iteratively propagating the tags of each photo to others and clamping the tags of exemplars. The process is illustrated in Algorithm 1.

Algorithm 1 Iterative Propagation Algorithm

Input: Similarity matrix $\boldsymbol{W}$, diagonal matrix $\boldsymbol{D}$ with $D_{ii} = \sum_j W_{ij}$.
Output: $\boldsymbol{Y}$.

1. Initialize the tag membership matrix $\boldsymbol{Y}$.
2. Update matrix $\mathbf{Y} = \boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{Y}$.
3. Clamp the tags of exemplars, i.e., let $\mathbf{y}_i = \mathbf{y}(x_i)$ if x_i is an exemplar, where $\mathbf{y}_i$ is the i-th row of $\boldsymbol{Y}$.
4. Repeat from step 2 until $\boldsymbol{Y}$ converges.

The process will converge to the solution of the following optimization problem [15].

$$\begin{aligned} &minimize \sum_{i,j=1}^{n} W_{ij}\|\mathbf{y}_i - \mathbf{y}_j\|^2 \\ &s.t. \mathbf{y}_i = \mathbf{y}(x_i) \quad \text{if } x_i \text{ is an exemplar.} \end{aligned} \quad (6)$$

Here we also adopt the similarity measure that explores both visual and temporal information (see Eq. 5), i.e., $W_{ij} = s(x_i, x_j)$. Based on the estimated $\boldsymbol{Y}$, we can easily obtain the binary tag membership by setting a threshold, i.e., $\mathbf{y}_k(x_i) = 1$ if the k-th entry in $\mathbf{y}(x_i)$ is above 0, and otherwise $\mathbf{y}_k(x_i) = 0$.

5. EMPIRICAL STUDY

5.1 Dataset and Feature Extraction

We conduct experiments with 10 different personal albums that are collected from Flickr. These photos are captured at different locations around the world and contain diverse content, including the records of cityscape, landscape, wide life, etc. Table 1 illustrates the number of photos in these albums. There are 4,576 photos in total.

Album	Photo Num.	Examplar Num.	%
Bombay	719	56	7.79
China	500	28	5.60
Korea	493	41	8.32
LongExposure	185	17	9.19
Manasquan	333	26	7.81
Nature	183	12	6.56
NewBook	157	19	12.10
Sunrise	286	21	7.34
Thailand	500	34	6.80
Wildlife	1,221	83	6.80
Average	458	34	7.42

Table 1: The numbers of photos and the selected exemplars for each album.

Many of the photos are with very high resolution. To speed up feature extraction, we scale the width of each photo to 240 pixels and then extract the following features: (1)225-dimensional block-wise color moment features generated from 5-by-5 fixed partition of the image; (2)128-dimensional wavelet texture features.

5.2 Empirical Results

We compare the following exemplar selection methods:

- Temporally consistent AP. The parameter σ_v is empirically set to the median value of the pairwise Euclidean distances of all samples and the parameter σ_t is empirically set to 1 hour (see Eq. 5). The parameter α is simply set to 0.5.
- AP, i.e., we employ the similarity estimated with only visual features.
- K-means clustering. For each cluster, the sample that is closest to the mean is selected as exemplar.
- Spectral clustering [16]. For each cluster, the sample that is closest to the mean is selected as exemplar.

In the last three methods, we set the number of clusters to be the same with the temporal consistent AP method (for AP method, we can tune the diagonal components of the similarity matrix to change the number of clusters [11]).

For tag inference, we compare two strategies, i.e., the naive tag assignment and tag propagation. Therefore, we will compare eight methods in all. The ground truths of the photos are established by ten volunteers as follows: for each album, the photos are exhaustively tagged by a volunteer. In this way, there are 6.93 ground truth tags associated with each photo in average.

For each photo, we estimate the precision, recall and F1-score measurements of the tags after performing the batch tagging approach. We average the F1-scores of all images and it is adopted as the performance evaluation measure of this work. Table 2 illustrates the performance comparison. From the results we can see that the AP method outperforms the K-means and spectral clustering methods with either naive tag assignment or tag propagation. In addition, the best result is obtained by the temporally consistent AP together with tag propagation, and this indicates the effectiveness of tag propagation and the integration of temporal

Method	Ave. F1-Score
Temporal Consistency AP+Tag Propagation	***0.7252***
AP+Tag Propagation	0.6604
K-Means+Tag Propagation	0.6244
Spectral Clustering+Tag Propagation	0.6494
Temporal Consistency AP+Naive Tag Assignment	0.6672
AP+Naive Tag Assignment	0.6374
K-Means+Naive Tag Assignment	0.5798
Spectral Clustering+Naive Tag Assignment	0.6227

Table 2: Performance comparison of different tagging methods.

information. Fig. 3 illustrates the detailed results, including average precision, average recall and average F1-score measurements for each album. The numbers of the selected exemplars are illustrated in Table 1. Comparing them with the sizes of the albums, the exemplars occupy only a very small portion (7.42% in average). Thus the smart batch tagging approach significantly reduces the human efforts in comparison with the exhaustive tagging, and fairly high tagging accuracy can still be maintained (average F1-score: 0.73).

For comparison, we also ask the volunteers to implement the naive batch tagging approach for each album, i.e., entering a set of tags for the whole album. Although this approach needs the least labor cost, the tagging accuracy is too low to be acceptable (the average F1-score is merely 0.45).

We have also conducted a user study with the ten volunteer users. They are asked to compare the smart batch tagging, exhaustive tagging and naive batch tagging schemes after experiencing all of them (we developed a tool that can support all the three schemes). Here we neglect the detailed numerical results due to the limitation of space, but the study results clearly indicate the superiority of the smart batch tagging scheme, and the ANOVA test shows that the difference is statistically significant.

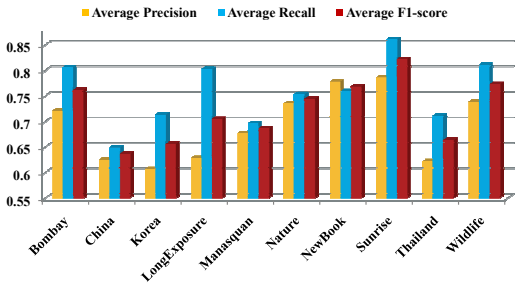

Figure 3: The performance evaluation measurements of smart batch tagging results.

5.3 Computational Cost

The computational cost of the batch tagging approach consists of the following four parts: (1) feature extraction; (2) similarity estimation; (3) temporally consistent AP-based exemplar selection; and (4) tag propagation.

In our practice, it costs about 0.062 second to extract features from each photo. For an album, the similarity estimation, exemplar selection and tag propagation averagely cost 1.7, 8.5 and 2.1 seconds, respectively. All these time costs are recorded on a PC with Pentium 3.40G CPU and 1G memory. We can see that the costs are fairly low and our user study results demonstrate that they are tolerable (in fact these time costs will be much less than the procedure of uploading photos if we apply the tool on photo sharing websites, and the feature extraction and similarity estimation can be implemented during the uploading process).

6. CONCLUSION

This paper describes a batch tagging tool that aims at helping users tag personal albums more efficiently. Instead of exhaustively tagging all photos or assigning tags to a whole album, the proposed approach automatically selects a set of representative exemplars from the album for manual tagging and the tags of the rest photos are automatically inferred. Empirical results on multiple albums have demonstrated the effectiveness of the proposed scheme.

There are several future works along this direction. One is to explore richer information sources that photos may contain, such as GPS coordinates [14]. Another work is to integrate tag recommendation to further facilitate the tagging process.

7. REFERENCES

[1] Flickr. *http://www.flickr.com.*

[2] J. Li and J. Z. Wang. Real-Time Computerized Annotation of Pictures. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol, 30, no. 6, 2008.

[3] J. Jeon, V. Lavrenko and R. Manmatha. Automatic Image Annotation and Retrieval using Cross-media Relevance Models. In *Proceedings of SIGIR*, 2003.

[4] J. M. Jia, N. H. Yu and X. S. Hua. Annotating Personal Albums via Web Mining. In *Proceedings of the 15th ACM International Conference on Multimedia*, 2008.

[5] R. Yan, A. Natsev and M. Campbell. A Learning-based Hybrid Tagging and Browsing Approach for Efficient Manual Image Annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[6] M. Ames and M. Naaman. Why We Tag: Motivations for Annotation in Mobile and Online Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing System*, 2007.

[7] L. S. Kennedy, S. F. Chang and I. V. Kozintsev. To Search or To Label? Predicting the Performance of Search-Based Automatic Image Classifiers. In *Proceedings of the 8th ACM International Workshop on Multimedia information retrieval*, 2006.

[8] D. Liu, X. S. Hua, L. J. Yang, M. Wang and H. J. Zhang. Tag Ranking. In *Proceedings of ACM International World Wide Web Conference*, 2009.

[9] B. Sigurbjörnsson and R. V. Zwol. Flickr Tag Recommendation based on Collective Knowledge. In *Proceedings of ACM International World Wide Web Conference*, 2008.

[10] H. M. Chen, M. H. Chang, P. C. Chang, M. C. Tien, W. H. Hsu and J. L. Wu. SheepDog-Group and Tag Recommendation for Flickr Photos by Automatic Search-based Learning. In *Proceedings of the 15th ACM International Conference on Multimedia*, 2008.

[11] B. J. Frey and D. Dueck. Clustering by Passing Messages between Data Points. In *Science*, vol. 315, 2007.

[12] Y. Jia, J. D. Wang, C. S. Zhang and X. S. Hua. Finding Image Exemplars using Fast Sparse Affinity Propagation. In *Proceedings of the 15th ACM International Conference on Multimedia*, 2008.

[13] W. T. Chu and C. H. Lin. Automatic Selection of Representative Photo and Smart Thumbnailing Using Near-duplicate Detection. In *Proceedings of the 15th ACM International Conference on Multimedia*, 2008.

[14] L. L. Cao, J. B. Luo and T. S. Huang. Annotating Photo Collections by Label Propagation according to Multiple Similarity Cues. In *Proceedings of the 15th ACM International Conference on Multimedia*, 2008.

[15] X. J. Zhu, Z. Ghahramani and J. Lafferty. Semi-supervised Learning using Gaussian Fields and Harmonic Functions. In *Proceedings of the International Conference on Machine Learning*, 2003.

[16] J. B. Shi and J. Malik. Normalized Cuts and Image Segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.22, no.8, 2000.