PROBABILISTIC MULTIMODALITY FUSION FOR EVENT BASED HOME PHOTO CLUSTERING

Tao Mei [†], *Bin Wang*[†], *Xian-Sheng Hua* [‡], *He-Qin Zhou* [†], *Shipeng Li* [‡]

[†] MOE-MS Key Laboratory of Multimedia Computing and Communication
University of Science and Technology of China, Hefei 230027, P. R. China
[‡] Microsoft Research Asia, 5F Sigma Center, 49 Zhichun Road, Beijing 100080, P. R. China

ABSTRACT

This paper presents a novel probabilistic approach to fusing multimodal metadata for event based home photo clustering. Photo events are characterized by the coherence of multimodality including time, content and camera settings. We incorporate these multimodal metadata into a unified probabilistic framework, in which event is taken as a latent semantic concept and discovered by fitting a generative model through an Expectation-Maximization (EM) algorithm. This approach is general and unsupervised, without any training procedure or predefined threshold. The experimental evaluations on $14k$ photos taken by 10 amateur photographers have indicated the effectiveness and efficiency of the proposed framework in browsing and searching personal photo collections.

1. INTRODUCTION

With the widespread adoption of digital cameras and cell phones used to capture people's memorable experience, the collections of digital photos are growing beyond the abilities of individual photographers to efficiently organize and search their photos. A natural approach to this issue is to temporally segment the photos into episodes or meaningful events, and sort both events and the photos within an event chronologically.

Automatic event or ontology detection of personal photos still remains a challenging issue, while most works in the literature reduce it to partitioning the photos' timestamp into contiguous segments that correspond to the underlying events. Typically, event is defined as the group of photos captured in relatively close proximity in time. Google Picasa [1] organizes the photos only by date information. Loui *et al.* [6] described a two class K-means based algorithm for event segmentation by using photo timestamps, and then broke an event into sub-events based on low-level content similarity. However, time and content information are treated separately in this work. PhotoTOC [7] used a locally adaptive threshold applied to time interval to group photos into table of contents. But the results are sensitive to the predefined thresholds, and content-based clustering is only used as a backup in post-processing as well. In [3], a local self-similarity based method was proposed for event clustering, in which temporal similarity was assessed at multiple scales. However, this work predominantly focused on time information, little content or other camera metadata were taken into account.

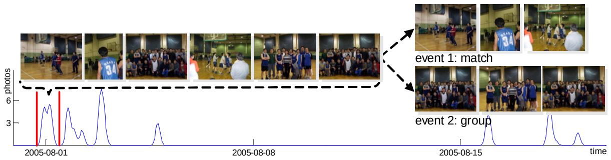

Fig. 1. The evolution curve of photo counts versus time. The horizontal axis is time, while the vertical is photo counts. The red vertical lines correspond to the event boundaries detected by typical time-similarity based approach.

It is observed that most of existing systems focused either on time or on content only, or used both but treated each in an independent way. However, a digital photo is usually recorded together with multimodal metadata such as image content (perceptual features) and contextual information (time and camera settings) [2][8]. An ideal solution to event clustering of personal photos is to automatically incorporate all these multimodal metadata into a unified framework, without being provided any *a prior* knowledge. Motivated by this point, we propose a novel approach to tackle this issue. Different from most of the previous work [6][7][3], in which event was defined only by time similarity, the photos are grouped into events in terms of both time and content coherence in our approach. In other words, the "event" in this paper can be deemed as "sub-event" or a definition in a finer granularity. Fig. 1 gives an example of the difference between the two types of event definitions. Although the six photos in this group were taken in a short period of time, they actually correspond to two events (match and group photos) rather than only one, in terms of content. Therefore, the multimodal information such as camera metadata and content are desirable

This work was performed when the first and second authors were research interns at Microsoft Research Asia. It is partially supported by the Science Research Fund of MOE-MS Key Laboratory of Multimedia Computing and Communication (Grant No. 05071805).

to be considered for event clustering.

The probabilistic framework for event based photo clustering in this paper is motivated by the modeling of latent semantic concept in [5][9]. The photo event is taken as a latent semantic concept, while the generation process of captured photos is modeled by a generative model. The multimodal metadata (time, content and camera parameters) of the photos belonging to the same event are assumed to exhibit coherence. These metadata are fused in a unified probabilistic framework, in which an EM algorithm is employed to estimate model parameters, and the number of events is determined by MDL (Minimum Description Length) principle.

2. PROBABILISTIC FRAMEWORK FOR EVENT CLUSTERING

According to the characteristics of photo events that both content and time of different photos appear similar within an event, the generation process of photos can be modeled by a generative model. The probabilistic framework fusing photos' multimodal metadata is described below.

2.1. Multimodal Metadata Analysis

The multimodal metadata used in our work include contextual information (time and camera parameters) and perceptual image content. The date/time stamp is recorded when a photo is taken by a digital camera or cell phone. Specifically, if EXIF [2] timestamp is not available, we rely on the creation time of image file instead. Camera parameters representing the capture condition can be also extracted from EXIF header. From the observation on a large photo collection, we utilize three event-related parameters (i.e. *aperture*, *exposure time* and *focal length*) instead of the seven as reported in [2]. Since these parameters have discrete values, they are quantized into several intervals. The perceptual content of a photo is described by three sets of most widely adopted features, i.e. *color*, *texture* and *face number*. We use color histogram (64D) and Tamura descriptor (20D) to represent *color* and *texture* features, respectively. To reduce the feature space and simplify the generative model described in Section 2.2, we perform PCA to select two subsets of features that mostly represent the *color* and *texture*, respectively, of the photo collection being investigated. We deal with *face number* in the same way as camera parameters, as it also has discrete values. As a result, a digital photo can be represented by a compact vector consisting of these multimodal metadata.

2.2. Generative Model of Photo Event

In the generative model, we assume all the photos belonging to a given photographer are known i.i.d. samples from an unknown distribution. Each photo $x_i \in X = \{x_1, x_2, \ldots, x_N\}$ corresponds to one unobserved semantic concept class – event $e_j \in E = \{e_1, e_2, \ldots, e_K\}$. N and K are the numbers of photos and events, respectively. As mentioned above, photo x_i can be represented by a triplet of observed multimodal metadata $(\mathcal{T}_i, \mathcal{C}_i, \mathcal{M}_i)$, where $\mathcal{T}_i$ is *time*, $\mathcal{C}_i$ is content feature list (i.e. *face number*, *color* and *texture*), and $\mathcal{M}_i$ is camera parameter list (i.e. *exposure time*, *aperture* and *focal length*). In order to simplify the model, the three types of modalities related to a photo x_i are assumed conditional independent given the respective hidden concept event e_j. Moreover, all the components of the different metadata belonging to a photo x_i are also assumed independent:

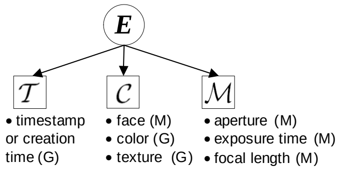

Fig. 2. Representation of the generative model of photo event. (M) – Multinomial distribution, (G) – Gaussian distribution.

$$p(x_i|e_j) = p(\mathcal{T}_i|e_j)p(\mathcal{C}_i|e_j)p(\mathcal{M}_i|e_j) = \prod_{\ell=1}^{L} p(x_{i,\ell}|e_j) \quad (1)$$

where $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,L})$, $x_{i,\ell}$ is the ℓth metadata of photo x_i $(\ell = 1, 2, \ldots, L)$, L is the size of metadata.

Each continuous component $x_{i,\ell}$ is modeled by single Gaussian distribution, while the discrete component $x_{i,\ell}$ is modeled by Multinomial distribution [1]. For example, since each event corresponds to a group of photos taken in a short period of time (refer to Fig. 1), *time* is modeled by single Gaussian distribution, in which the mean is the position of peak and the variance is the duration of the event; while *face number* and camera parameters follow Multinomial distributions.

As a result, the generating process of a photo x_i is described as follows:

1. Choose an event e_j with probability $p(e_j)$;
2. Generate a photo x_i with probability $p(x_i|e_j)$:
   - Select time $\mathcal{T}_i$ with single Gaussian distribution.
   - Select content feature $\mathcal{C}_i$. Each continuous component (i.e. *color* or *texture*) of $\mathcal{C}_i$ is generated by single Gaussian distribution, except that *face number* by Multinomial distribution.
   - Select camera parameter $\mathcal{M}_i$ with Multinomial distribution.

Fig. 2 shows the graphic representation of the generative model. To better understand the proposed model, we compare it to the model for News Event Detection [5]. In our

[1] Let Y denote the set of continuous metadata of a photo, Z denote the set of discrete metadata, we get Y=\{*color*, *texture*, *time*\}, Z=\{*aperture*, *exposure time*, *focal length*, *face number*\}.

model, a photo is equivalent to a news article; a photo group is equivalent to a news event; and a photographer is equivalent to a query keyword. Given a photographer, we are aiming at clustering all his personal photos into groups with each group corresponding to a latent event concept.

2.3. EM Learning

The model parameters can be estimated by maximizing the log-likelihood of the joint distribution:

$$\ell(X;\theta) \triangleq \log\big(\prod_{i=1}^{N} p(x_i|\theta)\big) = \sum_{i=1}^{N} \log\big(\sum_{j=1}^{K} p(e_j)p(x_i|e_j,\theta)\big) \tag{2}$$

where $p(x_i|e_j,\theta)$ is computed by (1) with the set of model parameters θ dropped. From the analysis of generation process, we derive that the model is a statistical mixture model which can be solved by applying EM algorithm as follows:

Step E

Compute the likelihood by (2)

Step M

I. Update posteriori of event e_j by Bayes rule:

$$p(e_j|x_i)^{new} = \frac{p(e_j)^{old}p(x_i|e_j)^{old}}{\sum_{j=1}^{K} p(e_j)^{old}p(x_i|e_j)^{old}}$$

II. Update model parameters of event e_j:

For each Gaussian metadata $x_{i,\ell}$ of photo x_i

$$\begin{aligned}\mu_{j,\ell}^{new} &= \frac{\sum_{i=1}^{N} p(e_j|x_i)^{new}\cdot x_{i,\ell}}{\sum_{i=1}^{N} p(e_j|x_i)^{new}}\\ \sigma_{j,\ell}^{new} &= \frac{\sum_{i=1}^{N} p(e_j|x_i)^{new}\cdot (x_{i,\ell}-\mu_{j,\ell}^{new})^2}{\sum_{i=1}^{N} p(e_j|x_i)^{new}}\end{aligned}$$

For each Multinomial metadata $x_{i,\ell}$ of photo x_i

$$p(x_{i,\ell}|e_j)^{new} = \prod_{m=1}^{M_\ell} [p(x_{i,\ell}=m|e_j)^{new}]^{\mathbf{1}(x_{i,\ell}=m)}$$

where M_ℓ is the number of discrete intervals for the ℓth metadata and $p(x_{i,\ell}=m|e_j)^{new} = \frac{1+\sum_{n=1}^{N} p(e_j|x_n)^{new}\cdot\mathbf{1}(x_{i,\ell}=m)}{M_\ell+\sum_{n=1}^{N} p(e_j|x_n)^{new}}$

III. Update model of event e_j:

$$p(e_j)^{new} \approx \frac{1}{N}\sum_{i=1}^{N} p(e_j|x_i)^{new}$$

$$p(x_i|e_j)^{new} = \prod_{\ell=1}^{L} p(x_{i,\ell}|e_j)^{new}$$

The E-step and M-step are iterated until convergence. The initial values of the parameters are computed using K-means. The number of events K must be determined in advance to perform EM iteration. Similar to [5][9], we adopt the MDL principle to select the best value of K:

$$K^* =^{argmax}_{K} \{2\cdot\ell(X;\theta) - m_K\cdot\log N\} \tag{3}$$

where $m_K = (K-1) + K\cdot N_G + K\cdot\sum_{\ell=1}^{N_M}(M_\ell - 1)$, N_G and N_M are the numbers of Gaussian and Multinomial distributions, respectively.

Table 1. Event clustering algorithm

```
Perform metadata extraction on photo collection
Initialize $K_{min}$ with the days of photo collection
for $K = K_{min}$ to $S \cdot K_{min}$ do
    Initialize model parameters $\theta$ by K-means
    for event $j = 1$ to $K$ do
        Update model parameters $\theta$ by EM
    end
end
Perform model selection by MDL (3)
Output $p(x_i|e_j)$, $p(e_j|x_i)$ corresponding to the best event number $K^*$
```

2.4. Event Clustering

The whole process for event clustering is summarized in Table 1, where S is the maximum number of events per day. With the proposed probabilistic model, we are able to concurrently obtain $p(x_i|e_j)$ and $p(e_j|x_i)$. The ith photo with the maximum *a priori* probability $p(x_i|e_j)$ among the photos in current event can be regarded as the representative photo, while the jth event with the maximum *a posteriori* probability $p(e_j|x_i)$ tells the corresponding event concept of photo x_i.

3. EXPERIMENTS AND EVALUATIONS

As aforementioned in Section 1, traditional photo event detection schemes reduced to detecting the boundaries in terms of timestamp. Thus the objective evaluations such as *precision* and *recall* can be obtained with the manually provided ground truth [6][3]. However, the photos within these detected boundaries usually do not correspond to a single event in our work due to the different definition of event concept. Therefore, we conducted three kinds of user studies and made comparison to the System Folder and PhotoTOC [7].

3.1. Data

We have collected $14k$ home photos taken by ten individuals in recent 3 years. These ten individuals are all amateur photographers having strong computer background. In our implementation, both the numbers of principle components for *color* and *texture* metadata are set to four. As a result, there are 13 dimensions of the multimodal metadata in total for each photo, including 9D Gaussian and 4D Multinomial distributions. Thus the parameters used for MDL in Eq. (3) are $N_G = 9$, $N_M = 4$, $M_\ell = 4$, $L = 13$. The data sets and event clustering results are listed in Table 2. It can be seen that our multimodal approach is able to detect more events (765 in total) than PhotoTOC (501). Accordingly, both time and content in an event represent more coherence, which makes the event concept more natural for organizing personal photos.

3.2. User Study

We evaluate our approach from three different viewpoints: (1) completion time for searching photos, (2) satisfaction score

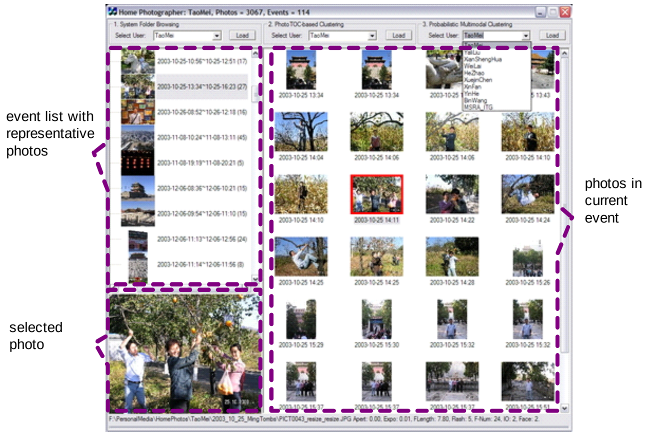
Home Photographer: TaoMei, Photos = 3067, Events = 114
1. System Folder Browsing
2. PhotoTOC-based Clustering
3. Probabilistic Multimodal Clustering
Select User:
TaoMei
Load
event list with representative photos
photos in current event
selected photo

Fig. 3. A prototype interface for event-based personal photo clustering and browsing. The representative photo for each event is bounded with red frame in the thumbnail view panel.

of the representative photos, and (3) satisfaction score of the overall usability. We also implemented System Folder and PhotoTOC for comparison. The user interfaces of the three schemes are designed analogously, as shown in Fig. 3. All of the ten individuals providing their personal photos are invited to do the user studies.

Completion time for searching photos The user study for the completion time is designed similar to the experiments in PhotoTOC. The three browsers were presented to each user in a random order, and the user did not know any background of the current browser. The scenario is that, given a set of sample photos randomly selected from his photo collection, the user tries to search the photo by using the three browsers respectively. There are totally 93 samples selected for the ten users. The average completion times (in seconds) for each photo are listed in the first row of Table 3. As we can see, by our multimodal approach, users can search photos more quickly than by the other two. The reason lies in that the multimodal approach is able to detect photo events not only by time but also by content. Thus an event usually contains more unitary perceptual content which facilitates users' searching.

Satisfaction score of representative photos and overall usability The second user study is to evaluate the representative photo for each event. After browsing all the photos in an event, the user was required to select an assessment for the representative photo: *poor*, *bad*, *neutral*, *good* or *excellent*. The satisfaction score of representativeness is defined as

$$\frac{N_{bad}\times 25+N_{neutral}\times 50+N_{good}\times 75+N_{exce.}\times 100}{N_{poor}+N_{bad}+N_{neutral}+N_{good}+N_{exce.}}$$

Finally, the user was invited to give an overall satisfaction score of the usability of browser. We only compare the two kinds of scores between PhotoTOC and our approach, since there is no general method for selecting representative photo in System Folder. From the last two rows of Table 3, we can see that our approach presents better results than PhotoTOC.

Table 2. Event clustering results

Person	Photos	K_1	K_2	Person	Photos	K_1	K_2
1	3067	139	266	6	720	56	52
2	642	48	54	7	1711	20	24
3	2903	71	101	8	217	13	35
4	2662	70	112	9	221	5	13
5	554	65	83	10	1101	14	25

Note: K_1 and K_2 is the number of events detected by PhotoTOC and our probabilistic multimodal approach, respectively.

Table 3. Evaluation of the three kinds of user studies

Browser	System Folder	PhotoTOC	Multimodal
Completion Time	36.55	30.43	21.90
Representativeness	–	51.65	69.09
Usability	–	60.25	77.75

4. CONCLUSIONS AND FUTURE WORK

In this paper, the problem of clustering personal photo collections into events is tackled by fusing the multimodal metadata in a unified probabilistic framework. We have shown that how the problem can be modeled using a generative procedure and then how the parameters in this model can be iteratively estimated using an EM algorithm. The proposed approach is fully automatic, unsupervised and extensible. It has many potential applications such as PhotoStory [5] and Photo Booklet [4]. In the future, we will incorporate more types of multimodal metadata and study their latent relationships. Furthermore, we will improve the representativeness of an event by considering more semantic features such as faces.

5. REFERENCES

[1] http://picasa.google.com/intl/zh-CN/.

[2] M. Boutell, J. Luo, "Bayesian Fusion of Camera Metadata Cue in Semantic Scene Classification," *In Proceedings of CVPR*, 2004.

[3] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal Event Clustering for Digital Photo Collections," *In Proceedings of ACM Multimedia*, 2003.

[4] X.-S. Hua, S. Li, and H.-J. Zhang, "Video Booklet," *In Proceedings of ICME*, 2005.

[5] Z. Li, B. Wang, M. Li, and W.-Y. Ma, "A Probabilistic Model for Retrospective News Event Detection," *In Proceedings of SIGIR*, 2005.

[6] A. C. Loui, A. E. Savakis, "Automatic Image Event Segmentation and Quality Screening for Albuming Applications," *In Proceedings of ICME*, 2000.

[7] J. C. Platt, M. Czerwinski, and B. A. Field, "PhotoTOC: Automatic Clustering for Browsing Personal Photographs," *Microsoft Research Technical Report MSR-TR-2002-17*, 2002.

[8] Y. Wu, E. Y. Chang, and B. L. Tseng, "Multimodal Metadata Fusion Using Causal Strength," *In Proceedings of ACM Multimedia*, 2005.

[9] R. Zhang, Z. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang, "A Probabilistic Semantic Model for Image Annotation and Multi-Modal Image Retrieval," *In Proceedings of ICCV*, 2005.