사진 사용 이력을 이용한 이벤트 클러스터링 알고리즘

김기응, 박태서, 박민규, 이영범, 김연배, 김상룡
삼성 종합기술원 Interaction Lab
{kekim, taesuh.park, minkyu03.park, leey, kimybae, srkim}@samsung.com

Adaptive Event Clustering for Personalized Photo Browsing

Kee-Eung Kim, Taesuh Park, Min-Kyu Park,
Yongbeom Lee, Yeunbae Kim, and Sangryong Kim
Interaction Lab, Samsung Advanced Institute of Technology

요 약

Since the introduction of digital camera to the mass market, the number of digital photos owned by an individual is growing at an alarming rate. This phenomenon naturally leads to the issues of difficulties while searching and browsing in the personal digital photo archive. Traditional approach typically involves content-based image retrieval using computer vision algorithms. However, due to the performance limitations of these algorithms, at least on the casual digital photos taken by non-professional photographers, more recent approaches are centered on time-based clustering algorithms, analyzing the shot times of photos. These time-based clustering algorithms are based on the insight that when these photos are clustered according to the shot-time similarity, we have "event clusters" that will help the user browse through her photo archive. It is also reported that one of the remaining problems with the time-based approach is that people perceive events in different scales. In this paper, we present an adaptive time-based clustering algorithm that exploits the usage history of digital photos in order to infer the user's preference on the event granularity. Experiments show significant performance improvements in the clustering accuracy.

Keyword: Information Retrieval (정보검색), Content/Applied Technology (콘텐츠/응용기술), Interface (인터페이스)

Introduction

The wide-spread use of digital cameras in everyday life has raised a new stance in information retrieval research. Traditional research approach to searching in the multi-media digital library has been focused on large amount of non-personal digital archive, often professionally organized and annotated. Hence, most of the effort has been geared toward advanced content analysis techniques such as computer vision, speech recognition, and natural language understanding. These techniques are not suitable for personal digital photos for a number of reasons. First, personal photos are rarely organized or annotated. Rodden and Wood [2003] report that most of the annotation activity is at most giving a name to the photo directory. Almost none of the subjects made annotation at the individual photo level. Second, even with advanced query interfaces such as QBIC [Flickner *et al.*, 1995] or keyword search, it is hard for the users to articulate what they are looking for. Such interfaces have not been quite successful since they ask for additional cognitive burden to the users. Third, these advanced techniques work reasonably well only under limited conditions, such as well-lighted and non-blurry images. Personal photos taken casually can rarely meet these conditions.

In contrast to the non-personal digital archive where most of the items are something that the user hasn't seen before, personal photos represent memory of events. It is also known that the chronological ordering of events is a dominant organization principle of human episodic memory [Tulving, 1983]. As such, most of the commercial digital photo management tools provide chronological view of photos as the primary browsing interface, and the recent focus on photo browsing interface has been centered on time-based clustering techniques to extract event boundaries from the photo archive. These algorithms generally analyze the time differences in photo shot times, and execute clustering algorithms to identify inter- and intra-event time intervals. Time-based clustering techniques for extracting "event clusters" have been suggested by Platt *et al.* [2003], Loui and Savakis [2000], and Cooper *et al.* [2003].

One remaining open question for the time-based clustering techniques is defining the granularity of events. Some users treat a multiple-day trip event as a single event, whereas others treat each day during the trip as separate individual events. Platt *et al.* [2003] also mentioned that finding the individual preference on the granularity of events is crucial for the time-based clustering algorithms. In this paper, we confirm this phenomenon, which will be discussed in the later sections, and present an algorithm to incorporate perceived differences in the granularity of events.

Time-Based Clustering Algorithms

In this section, we review some of the previous work on time-based clustering algorithms of digital photos. Particularly, we will see that these algorithms share the following common steps:

(1) Sort the digital photos in the chronological order by extracting shot times.

(2) Calculate the shot time intervals between subsequent photos.

(3) Compare the shot time intervals to determine whether the subsequent photos belong to the same event, or to different events. In general, if the time interval is significantly long, the two photos are determined to belong to different events.

For each algorithm, we will identify the key equations that determine whether inter- or intra-event shot time intervals. These equations will serve as the basis for extending the algorithms to incorporate usage histories of photos and having better results on event clustering.

Loui and Savakis [2000]: Let g_i be the shot time difference between the i-th photo and $i+1$-th photo when sorted in the chronological order. The algorithm takes the histogram of g_i's and performs the 2-means clustering. The cluster with the smaller centroid represents the shot time differences of intra-event photos, whereas the other cluster represents that of inter-event photos. Hence, the key equation for deciding whether g_i corresponds to the inter-event photo interval is given by

$$p_1(g_i) < p_2(g_i), \qquad (1)$$

where $p_1(g_i)$ denotes the probability of g_i belonging to the cluster with the smaller centroid, and $p_2(g_i)$ denotes the probability belonging to the one with the larger centroid. If the above equation holds true, then the algorithm splits between the i-th and $i+1$-th photos, and decides them to be in different events.

Platt *et al.* [2003]: The algorithm takes a similar approach, but uses different formula for deciding intra- or inter-events. Specifically, the formula is given by

$$\log(g_i) > K + \frac{1}{2d+1}\sum_{j=-d}^{d}\log(g_{i+j}), \qquad (2)$$

where K is the constant chosen to be $\log(17)$, and d is the window size chosen to be 10. This equation essentially compares g_i to the local geometric average of $(2d+1)$ time differences, and decides g_i to be inter-event when g_i is large enough compared to the average.

Cooper *et al.* [2003]: The original algorithms presented in the paper have various versions, including the algorithm that considers both the time *and* content differences. However, for the sake of brevity, we summarize the time-based only version of the algorithm.

First, the time similarities are calculated for each and every pair of photos (not just subsequent pairs),

$$S_T(i,j) = \exp\left(-\frac{|t_i - t_j|}{T}\right),$$

where t_i is the shot time of the i-th photo, and T is the time scale of either 10^3, 10^4, 10^5 minutes. Increasing the time scale T will show a coarser clustering of photos.

To calculate the cluster boundaries, Cooper *et al.* define the following novelty score for each photo

$$\upsilon_T(i) = \sum_{k,k'=-L}^{L} S_T(i+k, i+k')G(k,k'),$$

where G is the Gaussian kernel with width 12, so that $L = 6$. The boundaries are defined as the local peaks in the novelty scores. The algorithm identifies the boundaries for each time scale value, and then selects the best time scale based on the confidence score defined as,

$$C(B_T) = \sum_{l=1}^{|B_T|-1} \sum_{i,j=b_l}^{b_{l+1}} \frac{S_T(i,j)}{(b_{l+1}-b_l)^2} - \sum_{l=1}^{|B_T|-2} \sum_{i=b_l}^{b_{l+1}} \sum_{j=b_{l+1}}^{b_{l+2}} \frac{S_T(i,j)}{(b_{l+1}-b_l)(b_{l+2}-b_{l+1})},$$

where $B_T = \{b_l\}$ is the set of identified boundary photos for the time scale T. The final set of boundary photos is the B_T that maximizes $C(B_T)$. In short, the first term is the average intra-cluster similarity between the photos, and the second term is the average inter-cluster similarity between photos in adjacent clusters. Detailed discussions on the novelty score and the confidence score are beyond the scope of this paper, but they can be found in the original paper.

Quantifying the Photo Usage

Although the time-based clustering algorithms summarized in the previous section are effective in practice, there is still one remaining issue – differences in the perceived time-scale granularity of events. According to our initial experiments involving three end-users, we discovered that the preference on the granularity of events can differ by as large as a factor of 10000. Cooper *et al.* [2003] partly address this issue by calculating clusters at various time scales, and choosing the best time scale. However, we suspect that we can gather more useful information about the user's preference on the granularity of events by observing the usage behavior of her personal digital photos. Such usage behavior includes the selection for blogging or e-mailing with other photos, or the difference in the keyword tags if any. Particularly, the algorithm to be presented in the next section will make use of the following two types of photo usage behaviors. (Figure 1)

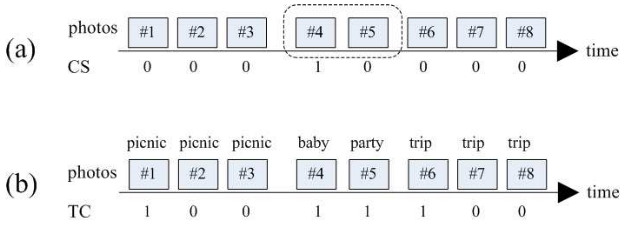

Figure 1. Illustrations of (a) Cross Selection and (b) Tagset Change of photos in chronological order.

Cross Selection $CS(i)$ takes the value 1 if and only if $\exists k \le i, k' \ge i+1$ such that k-th photo and k'-th photo have been selected together for blogging, e-mailing, etc. The idea here is that users are *likely* to handle intra-event photos together. When writing about something using more than one personal photo, we suspect that the unit of the article will be most likely an event and that detecting these selections will provide a useful hint about event boundaries.

Tagset Change $TC(i)$ takes the value 1 if at least one of the i-th and $i+1$-th photos are tagged (or in other words labeled) by the user and the tags differ. $TC(i)$ takes the value 0 if at least one of the i-th and $i+1$-th photos are tagged (or in other words labeled) by the user and the tags are the same. $TC(i)$ takes the value 0.5 if neither of the photos are tagged. The idea here is that the users typically tag their photos per event basis, and the users hardly tag their photos individually. This behavior is also reported in Rodden and Wood [2003]. As such, the difference in the tag set between subsequent photos provides a useful hint about event boundaries.

Note that the functions $CS(i)$ and $TC(i)$ are the features of the photo usage, reflecting a small aspect of personal digital photo lifecycle. The algorithm that we will present in the next section is not necessarily confined to these two features. As we understand how the user creates, manages, and recycles her personal digital photos, we can extract new features and extend the algorithm.

Adaptive Time-Based Clustering Algorithm

A crucial step in making the algorithms adaptive is identifying the event boundary decision criteria in the previous non-adaptive time-based clustering algorithms, and incorporating the photo usage features into the decision criteria. In this section, we will show how we can extend the non-adaptive time-based clustering algorithms case by case. However, the extension technique doesn't necessarily confine to the non-adaptive time-based clustering algorithms shown here, and other algorithms can be similarly made adaptive.

Adaptive version of Loui and Savakis [2000]: The decision criteria shown in Equation 1 can be transformed to a sigmoid function

$$\frac{1}{1+e^{-x}} > 0.5 \qquad (3)$$

where

$$x = -p_1(g_i) + p_2(g_i)\,.$$

This is actually a one-layer Perceptron with two input units and one sigmoid output unit. We generalize this Perceptron output unit to have input nodes for $CS(i)$ and $TC(i)$, and we have the new definition

$$\begin{aligned} x &= w_1 p_1(g_i) + w_2 p_2(g_i) \\ &\quad + w_3 CS(i) + w_4 TC(i) + w_5 \end{aligned}.$$

The optimal values for the parameters (or the weights in the Perceptron) w_1, w_2, w_3, w_4 and w_5 are calculated via standard Perceptron learning rule [Duda *et al.*, 2000].

In order to apply the Perceptron learning rule, we need some sort of training data of event boundaries. The training data can be gathered through a user interface that lets users to move photos between event clusters. Note that as more training data are gathered, repeated re-training of the parameters will result in optimal values for each particular user, hence we obtain the personalized model of event granularity.

Adaptive version of Platt *et al.* [2003]: The decision criteria shown in Equation 2 can be transformed to the same sigmoid function (Equation 3), but with

$$x = \log(g_i) - K - \frac{1}{2d+1}\sum_{j=-d}^{d}\log(g_{i+j})\,.$$

We follow the same technique as in the previous case by generalizing the sigmoid function to

$$\begin{aligned} x &= w_1 \log(g_i) + w_2 \frac{1}{2d+1}\sum_{j=-d}^{d}\log(g_{i+j}) \\ &\quad + w_3 CS(i) + w_4 TC(i) + w_5 \end{aligned}.$$

Note that the constant term K has disappeared in the above formula, since the bias term w_5 can reflect the value. The optimal values for the parameters w's are calculated via Perceptron learning rule, same as the previous case.

Adaptive version of Cooper *et al.* [2003]: Similarly, the sigmoid function representation of the decision criteria becomes

$$x = f(\upsilon_T(i-1), \upsilon_T(i), \upsilon_T(i+1))$$

where the function $f(\upsilon_T(i-1), \upsilon_T(i), \upsilon_T(i+1))$ yields value 1 if $\upsilon_T(i)$ is a local maxima, and -1 if not. Hence the generalized sigmoid function for incorporating the photo usage would be

$$\begin{aligned} x &= w_1 f(\upsilon_T(i-1), \upsilon_T(i), \upsilon_T(i+1)) \\ &\quad + w_2 CS(i) + w_3 TC(i) + w_4 \end{aligned},$$

where the parameters w's are calculated via Perceptron learning rule, same as the previous case.

Preliminary Experiments

In order to compare the performances of the algorithms, we collected personal photos from 3 users, user #1, user #2 and user #3. Specifically, we gathered 716 photos spanning 1012 days from user #1, 1204 photos spanning 1539 days from user #2, and 207 photos spanning 509 days from user #3. These photos were loaded into commercial off-the-shelf photo management software in order to receive information about event boundaries directly from the users. This information was then used to calculate optimal parameters in the adaptive time-based clustering algorithm. In our case, we used one-layer Perceptron for deciding the event boundaries.

After the optimal parameters are found, we tested the algorithm on the same dataset in order to calculate recall, precision, and F-measure. These performance measures are calculated as follows: The recall is defined as the ratio of the number of correct event boundaries found by the algorithm to the number of event boundaries specified by the user. The precision is defined as the ratio of the number of correct event boundaries found by the algorithm to the number of total event boundaries yielded by the algorithm. The F-measure is the geometric average of recall and precision. In the extreme case, if the algorithm identifies every photo as an event boundary, the precision would be 1.0 but the recall would be very low. If the algorithm identifies the whole photo archive as a single event, the recall would be 1.0 but the precision would be very low. Hence, the F-measure is widely used in order to measure the performance more accurately.

Note that even though optimal parameters are found, the adaptive time-based clustering algorithm does not always show the perfect performance, *i.e.*, 100% accuracy in finding the event boundaries. This is because there may be some data points (event boundaries) that are beyond the representation of the decision model. Even though we used the simplest, one-layer Perceptron for modeling event boundaries, our algorithm shows superior performance over those from non-adaptive algorithms. We compare the performances among two non-adaptive clustering algorithms, the Platt *et al.* and the Loui *et al.*, and the adaptive version of Platt *et al.* algorithm. The table below summarizes the experimental results.

User	Algorithms	Recall	Precision	F-Measure
User #1	Platt *et al.*	1.0	0.47	0.64
	Loui *et al.*	1.0	1.0	1.0
	Adaptive	1.0	1.0	1.0
User #2	Platt *et al.*	0.91	0.43	0.59
	Loui *et al.*	1.0	0.21	0.35
	Adaptive	0.78	0.80	0.79
User #3	Platt *et al.*	0.86	0.83	0.85
	Loui *et al.*	0.97	0.58	0.73
	Adaptive	0.93	0.93	0.93

As we can see, the adaptive version of the Platt *et al.* algorithm out-performs other non-adaptive algorithms in all photo archives. This result is expected since the adaptive algorithm is an extension to the non-adaptive algorithm, and hence, the adaptive algorithm will show the same accuracy as that of the non-adaptive algorithm in the worst case. The result above is preliminary in the sense that we are implementing the adaptive versions of other time-based clustering algorithms. However, we expect that we will achieve similar level of improvement in other adaptive algorithms.

Conclusions and Future Work

In this paper, we have presented an adaptive algorithm for clustering personal digital photos in the unit of events. When clustering the photos according to events, it is crucial to know the duration of events preferred by the user. The algorithm is adaptive in the sense that it will infer the event duration preference from the photo usage history. We described how the existing non-adaptive algorithms can be extended to be adaptive, and this technique can be applied to a wide variety of non-adaptive time-based clustering algorithms. Experiments show that the adaptive algorithm produces far better results compared to the non-adaptive algorithms. Currently, we are developing personal digital photo management software with the adaptive clustering algorithm.

As for the direction of future research, we are investigating other photo usage features that may provide hints about perceived granularity of events, as well as applying the extension framework to other non-adaptive time-based clustering algorithms. We are also looking into the ways for summarizing events – selecting representative photos, or providing with extra information that helps users to conjure memory.

References

E. Tulving (1983). *Elements of Episodic Memory*. Oxford University Press, 1983.

M. Flickner, D. Petkovic, D. Steele, P. Yanker, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner and D. Lee (1995). Query by image and video content: The QBIC system. In *IEEE Computer*, 28 (9).

R. Duda, P. Hart, and D. Stork (2000). *Pattern Classification*. Wiley-Interscience, 2000.

A. Loui and A. Savakis (2000). Automatic image event segmentation and quality screening for albuming applications. In *Proceedings of IEEE International Conference on Multimedia and Expo 2000*.

M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox (2003). Temporal event clustering for digital photo collections. In *Proceedings of ACM International Conference on Multimedia 2003*.

J. Platt, M. Czerwinski, and B. Field (2003). PhotoTOC: Automatic clustering for browsing personal photographs. *Technical Report MSR-TR-2002-17*, Microsoft Research, 2002.

K. Rodden and K. Wood (2003). How do people manage their digital photographs? In *Proceedings of CHI 2003*.