Home Photo Retrieval: Time Matters

Philippe Mulhem, Joo-Hwee Lim

Home Photo Retrieval: Time matters

Philippe Mulhem[1] and Joo-Hwee Lim[2]

1. IPAL-CNRS, 2. Institute for Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613, Singapore.
Email: {mulhem,joohwee}@i2r.a-star.edu.sg

Abstract

Temporal information has been regarded as a key vehicle for sorting and grouping home photos into albums associated with events. While time-based browsing might be adequate for relatively small photo collection, query and retrieval would be very useful to find relevant photos of an event in large collection. In this paper, we propose the use of temporal events for organizing and representing home photos using structured document formalism and hence a new way to retrieve photos of an event using both image content and temporal context. We describe a hierarchical model of temporal events and the algorithm to construct it from a collection of home photos. In particular, we compute metadata of a node from the metadata of its children recursively to facilitate content-based and context-based matching between a query and an event. With semantic content representation extracted using Visual Keywords and Extended Conceptual Graphs, we demonstrate the effectiveness of photo retrieval on 2400 time-stamped heterogeneous home photos with very promising results.

1 Introduction

It had been shown [15,16] that time is the most prominent aspect used by consumers to retrieve their photos: when they want to retrieve photos of one event, they consider "*I roughly know when it is [...]*" (p. 5 of [16]) and they browse their collections until they found what they are looking for. Indeed temporal information has been utilized for sorting and grouping home photos into albums associated with events, which is the most requested feature for photo organization [15].

For instance, MyPhotos [18] proposes effective temporal navigation in photo collections. The Hierarchical Browser described in [7] exploited photo creation time to cluster photos and to generate meaningful summaries. The PhotoTOC [13] and AutoAlbum [12] used adaptive gap threshold to split photos sorted by creation time into albums and a left-right Hidden Markov Model on the color information to further cluster the ordered photos in large albums. These works share a common insight and assumption: the irregular and recursive

busty time patterns of home photos taking correlate well with events [7] and a noticeable time gap correspond to a change in event [13].

While time-based browsing might be adequate for relatively small photo collection (average of 1000 photos in [16]; 1300 or less photos in [13]), query and retrieval would be very useful to find relevant photos of an event in large collection. In this paper, we propose the use of temporal events for organizing and representing home photos using structured document formalism and hence a new way to retrieve photos of an event using both image content and temporal context.

Though time-based photo clusters are good and sensible approximation to photo events, we remind that in some situations, the time gap may not reflect the true semantic boundary of an event. Hence we prefer to use the term "temporal event" to reflect the approximation in this paper. Furthermore, we draw the analogy of a home photo collection to a structured document. Images grouped into temporal events are indeed considered similar in some way to parts in a structured document as shown in Fig. 1. The images at the leaf nodes correspond to passages in a structured document.

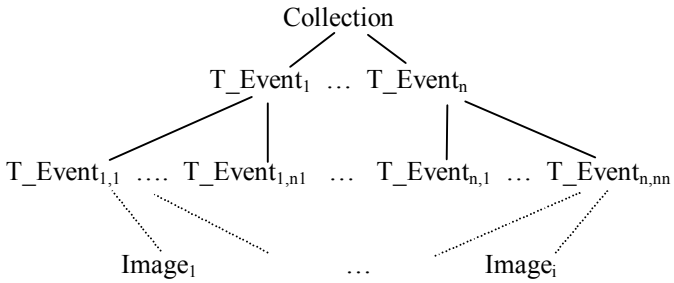

Fig. 1. A typical temporal organization of a home photo collection

For such documents, several approaches try to use the explicit structure to provide clues to retrieve the best document parts that match the queries: [2] for the FERMI Esprit II project defined a logical-based expression of information propagation and query processing. This work was used in [8] to provide a Dempster-Shafer frameword for structured document retrieval. Other approaches use the structure to facilitate the computation of relevance status values, but not for the selection of the level of structure retrieved: the work reported in [19] employed the usual vector space model to index whole documents and document parts and to retrieve document parts, showing that the context of document parts may be used to retrieve relevant chapters of legal documents; [11] use probabilistic belief networks to represent the impact of retrieval of the structure of documents. Our approach is more related to the latter approaches: we study the impact of the temporal structure on image retrieval.

From another point of view, some works have considered different uses of links in, or between, web or hypertext pages. [9] defined a notion of "local neighborhood" of a web page that is used to compute the relevance of a page using the out-going links to provide better results than usual non-contextual searches. [4] shown that the links between web pages are important to find out what are the "best" entries of the entire web sites. [6] used the links of texts nodes connected to image nodes in hypertext documents to index images. [1] uses the context of occurrence of images or video links to classify them. We adopt similar belief that

the composition of temporal events may be used as contexts during the query processing of image retrieval.

We organize the rest of the paper as follows. We describe a hierarchical model of temporal events and the algorithm to construct it from a collection of home photos in the next section. The metadata definition of a temporal event based on the metadata of its children as well as content-based and context-based query processing is presented in Section 3. We demonstrate the effectiveness of photo retrieval on 2400 time-stamped heterogeneous home photos with very promising results in Section 4 followed by conclusion.

2 Temporal Events Modeling

A temporal event $T \in \mathbf{T}$, on a collection of image $\mathbf{I}$ is a triplet (*Sub-T, Image, Index*) where *Sub-T* ($\subset \mathbf{T}$) is the set of temporal events that compose *T*, *Image* ($\subset \mathbf{I}$) is the set of images that compose *T*, and *Index* is the content representation of *T*. We ensure that a temporal event *T* is composed of either images or sub-events but not both i.e. $T.Sub\text{-}T \ XOR \ T.Image = \varnothing$ where a dot notation *A.B* indicates the access to the element *B* of the instance *A*.

An image *Im* is described by a triplet (*Image_Data, Time, Index*), with *Image_Data* being the raw pixel data of the image, *Time* as time and date of image creation, *Index* as the metadata describing the content of the image. We define F_{images} as a mapping from $\mathbf{T}$ to $\mathbf{I}$ that contains the list of all the images (directly or transitively) belonged to a temporal event. Using the F_{images} function, we formalize the fact that a temporal event contains only consecutive images:

$$\forall I_1, I_2 \in F_{images}(T), \neg(\exists I_3 \in \mathrm{I} \setminus F_{images}(T) \ \text{between}(I_1.Time, I_2.Time, I_3.Time))$$

where *between* denotes a ternary predicate that is true if the time symbol in the third place is between the other two time symbols.

To build a hierarchy of temporal events, we use an approach inspired from the single-link clustering [5, p.233]. For level of granularity G_k, we compute the matrix storing the image×image temporal differences. Then we use a one-pass single link algorithm to build the temporal events using these temporal distances and a threshold over which the images are not linked. We denote I_{Gk} the set of initial temporal events for the granularity G_k. A nice property of this approach is that for two granularities G_i and G_j with respective thresholds Th_{gi} and Th_{gj}, $Th_{gi} < Th_{gj}$, we ensure that for each initial temporal event T_l of T_{Gi} the images of T_l are strictly included in the images of one and only one initial temporal event T_m of T_{Gj}. This is very similar to that of PhotoTOC [13] and the first step of the clustering in [7]. In implementation, the difference between Gregorian date and time are computed with the use of Julian day numbers. For example, consider a collection *I* of 6 images with a DDMMYYYY time format, I = {Im_1[15081998], Im_2 [16081998], Im_3[19081998], Im_4[20081998], Im_5[27081998], Im_6[28081998]}. A first generation of initial temporal events T_{G1} for a threshold of one day is {{Im_1, Im_2}, { Im_3, Im_4}, {Im_5, Im_6}}. The generation T_{G5} of a threshold of 5 days is {{Im_1, Im_2, Im_3, Im_4}, {Im_5, Im_6}}. In this paper, we decided to limit the temporal definition using temporal features, to be able in the future do study the impact of other ways (like imag content) to define such events.

3 Content Representation and Query Processing

The content representation (or index) of a temporal event is based on the content representation of its composing images. For a temporal event T, we define $T.Index = F_{index}(\{im.time,im.index\} \mid im \in F_{images}(T)\})$ where the function F_{index} computes the index of a temporal event based on the time and indexes of the images that belong to it. In this paper, the index associated with an image I, denoted as ($I.Index_{vk}$, $I.Index_{cg}$), is composed of both a Visual Keywords representation $Index_{vk}$ and an Extended Conceptual Graph representation $Index_{cg}$ (please refer to [10] for details). We compare in this paper two simple definitions of the function F_{index}:

- The index of a temporal event is that of the image with the highest burst rate in the temporal event suggesting that this photo is important (as mentioned in [16], people take several digital photos to ensure they obtain a good one);
- The index of the temporal event is considered as a simple concatenation of the indexes of the images that belong to the event;

3.1 Image Index based on Visual Keywords

Visual Keywords are local semantic regions derived from statistical learning (e.g. Support Vector Machines). Fig. 2 shows the 26 classes of Visual Keywords defined and learned for the 2400 home photos in our experiments. For the purpose of image indexing, multi-scale view-based recognition against these 26 Visual Keywords are performed on each image and the probabilistic recognition results are reconciled upon multiple resolutions and aggregated according to configurable spatial tessellation (e.g. five coarse areas: left, right, top, bottom, center) as the image index. In essence, an image area is represented as a histogram of Visual Keywords based on the certainties of local recognition. The similarity matching function beween a query Q_{vk} represented by the index of a query image and the index of a image I of the database, $Match_{VK}(Q_{vk}, I.Index_{vk})$, is simply a weighted similarity between the corresponding areas in the query and the image (for example, we can assign higher weight for the center area as consumers tend to place the subject at the center while taking pictures). The similarity between two images areas is a simple histogram city-block distance computed between the visual keywords histograms in which the bins now possess semantic meanings. The choice of visual keywords is specific to a given context, because it's well known that the current state of the art of computer vision and image understanding is far from being able to provide accurate results in any context. So, the Visual Keywords are defined *a priori* and each Visual Keyword specificity is learned using examples, and each photo of the collection is then labelled automatically without any human intervention.

Fig. 2. The 26 Visual Keywords adopted for home photos.

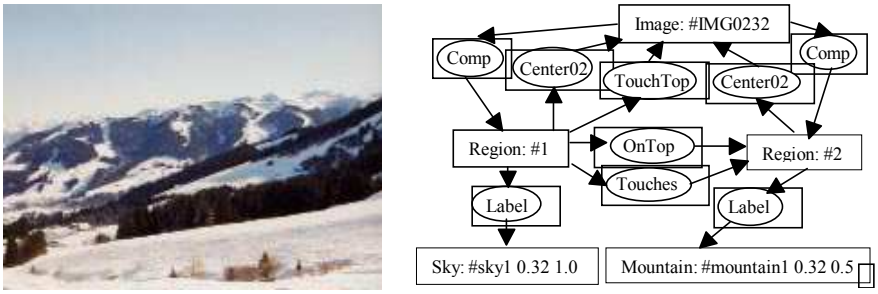

Fig. 3. The conceptual graph of an image.

3.2 Image Index based on Extended Conceptual Graphs

The description of the images using Extended Conceptual Graphs is achieved through the use of a knowledge representation formalism, namely the Conceptual Graphs [17]. Because of space constraints, we only present an example in Fig. 3, where the left part presents a photograph and the right part an excerpt of the conceptual graph that index this image. Such conceptual graph representation is based on the VK labeling process decribed above, but the advantage of using graphs if the possiblity of simulating deduction. Compared to the initial definition of conceptual graphs by Sowa, we use additional elements dedicated to represent the certainty of recognition of elements (for instance, the Mountain concept of Fig. 3 is recognized with a certainty of 0.5), and the importance of the elements (for instance the Sky concept of Fig. 3 has an importance of 0.32). The query processing for the conceptual graphs is related to the function $Match_{CG}(Q_{cg}, I.Index_{cg})$, and based on sub-graph matching (the query Q_{cg} being also a conceptual graph), considering the importance and certainty of recognition of the elements of the image index $I.Index_{cg}$ and the query. The query Q_{cg} may be an image index, as in the experiments conducted in Section 4. The matching value uses the concepts weights as well as relationships weights.

3.3 Selection-based Temporal Event Index

The first option, in order to generate the index of a temporal event, is to select one image, hopefully the most representative, and to use its index as the index of the temporal event.

To integrate the use of such temporal aspects when finding the most important image of a temporal event, we define the function f_k, from N^+ (going from the time and date of the first image of the event, in seconds, to the time and date of the last image of the event) to N^+ (image numbers going from *1* to *y*), as an interpolation of the cumulative function indicating for one time point t_x the number of images taken before (or at) t_x in the event. Such a function f_k is generated by using cubic spline interpolation [14]. The temporal representative value of an image *p* is then computed as the absolute value of the derivative $f'_k=df k/dt$ for the value t_p.

The index of a temporal event T is then similar to that of an image i.e. ($T.Index_{vk}$, $T.Index_{cg}$). The query processing is computed using a matching function $Match_{TE}$ defined as a linear combination of the $Match_{cg}$ and $Match_{vk}$ of the query and the representative image:

$$Match_{TE}(Q,T.Index) = \alpha*Match(Q_{vk},T.Index_{vk}) + (1- \alpha)*Match(Q_{cg},T.Index_{cg})$$

3.4 Aggregation-based Temporal Event Index

If one is not convinced that a selected image is always adequate to represent a temporal event effectively for different queries, then as a second approach we represent the index of a temporal event as the union of the indexes of its images, so *T.Index* is a set of couples ($c.Index_{vk}$, $c.Index_{cg}$), one c per image of *T*. Query processing is based on the maximum matching (i.e. best match) obtained between the query and the indexes:

$$Match_{TE}(Q,T.Index) = \max_{c \in T.Index} \{\alpha*Match(Q_{vk},c.Index_{vk}) + (1- \alpha)*Match(Q_{cg},c.Index_{cg})\}$$

3.5 Query Processing for Temporal Events

The query processing for a temporal event is analogous to that for text passage retrieval [19] with a query image being a passage and a temporal event as a document part. The retrieval function performed for a query *Q* and one image *I* depends on 3 representations: the query representation, the index of *I*, and the indexes of the temporal events that contain *I*. As we described earlier, in our case the query *Q*, the image index, and temporal event index are based on both VK and CG representations.

Using the temporal event indexes during query processing, we intend to capture both the relevance of an image *per se* and the relevance of temporal context. We assign the importance of a temporal event as inversely proportional to its duration for query processing. Hence we compute the relevance status value of an image *Im* for a query *Q* as

$$RSV(Q,I) = \alpha*Match_{vk}(Q_{vk}, I.Index_{vk}) + (1- \alpha)*Match_{cg}(Q_{cg},I.Index_{cg}) + \sum_{T \in \mathbf{T} \text{ s.t. } I \in F_{images}(T)} \beta(T,I) * Match_{TE}(Q,T.Index)$$

where β is a function that returns the value of importance of a temporal event based on its duration and on the structural distance between the temporal event *T* and *I*, assuming that the longer an event is, the lower it representents accurately one image, and that the further the event from the image in the structure representation the lower in impacts the image retrieval.

4 Experiments on Image Retrieval by Content and Context

The experiments were conducted on a set I of 2400 home photos with temporal metadata year, month, day, hour, minute, and second. The temporal events were built as described in Section 2, with a granularity level of 1 hour: photos taken within an one-hour interval are assumed to be related to a short event. Thus in this paper, we only consider one level of temporal events that plays important role as temporal context. The number of temporal events obtained is 278, giving an average of 8 photos per event, with a maximum of 90 images and a minimum of 1 image. The β function is a linear function that only considers 1) temporal events duration within a day (i.e. 86400 seconds), assuming the existence of the function *duration(T)* that gives the duration (in seconds) of the event *T*, and 2) the inverse of the distance (*Str_distance*) between the event and one image in the events structure:

$$\beta(T,I) = \begin{cases} \frac{1}{Str\_distance(T,I)} \cdot (1 - \frac{duration(T)}{86400}) & \text{if } duration(T) \le 86400 \\ 0 & \text{if } duration(T) > 86400 \end{cases}$$

The experiments were conducted on a set of 26 queries, using query by example, with the VK (c.f. 3.1) and CG (c.f. 3.2) representations, using α equal to 0.9. The assesment of the relevant images for each query has been made by two persons, and we kept the intersection of the two assessments. The queries correspond to different runs of the system to retrieve images corresponding to meals, weddings, parks, watersides and beaches, swimming pool, streets and roadsides photos. The images of the query by examples were selected randomly from the relevant set for each query. We present the recall versus precision[1] curves (Fig. 4), averaged over the 26 queries, obtained with and without the use of the temporal events using our approach (denoted as SYMB) and we compare as a baseline to a color-only approach using a 4x4 grid of HSV color-space local histograms (similar to the PicHunter System [3], denoted as HSV). We study the use of the representative image (c.f. 3.3, denoted as Rep) and the use of the whole set of image indexes (c.f. 3.4, denoted as Max).

In Fig. 4, we see that the use of both representative-based or max-based temporal events increases the quality of the results obtained: +5.6% for Rep SYMB compared to SYMB, +6.9% for Rep HSV, but the most convincing results are +25% for Max SYMB as well as for Max HSV. This clearly shows that the use of contextual information for the retrieval of home photos impacts positively each of the content representations. It also shows that the use of the representative image (Rep) does not perform as well as the Max, as expected. We also

[1] Using the trec_eval software from ftp://ftp.cs.cornell.edu/pub/smart/ .

confirm that our use of symbolic representations surpasses the low-level color-based representation scheme. We notice also that the Max HSV scheme (average precision 0.3178) performs almost as good as the basic SYMB scheme (average precision of 0.3253), wich reinforce our idea that the context is important when considering image retrieval.

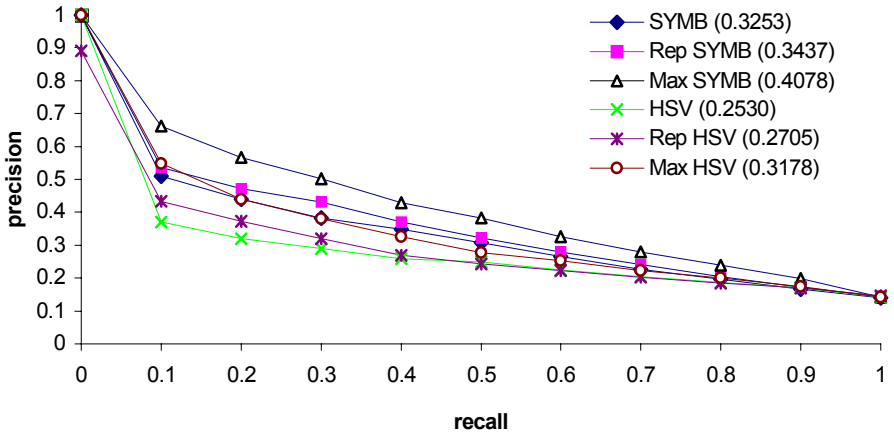

Fig. 4. Recall Precision with and without using the temporal events.

Table 1 presents average precision values at top 5, 10, 20 and 30 photos for the 26 queries. The values in parentheses indicate the relative increase. This table highlights the importance of the precision improvement when presenting results to the user. For instance, at top 10 photos, the average number of relevant photos is 0.65 for SYMB and 0.8 for Max SYMB (meaning that there is on average 8 relevant photos among the first 10 retrieved), and 0.4 for HSV and 0.65 (on average 6.5 relevant photos in the first 10 retrieved). At top 30 photos, the number of relevant documents is on average 15 for SYMB and 20 for Max Symb, and also 10 for HSV and 15.9 for Max HSV; the precision of Max SYMB at top 30 photos is greater than 2/3, empirical threshold under which we consider such system to be unusable by consumers. We notice once again that the Max HSV using the event infomartion performs as well as the basic SYMB sheme.

Table 1. Average precision values at 5, 10, 20 and 30 images.

Avg. Prec.	SYMB	Max SYMB	HSV	Max HSV
@ 5 photos	0.731	0.88 (+20.4%)	0.508	0.72 (+41.8%)
@ 10 photos	0.658	0.80 (+21.6%)	0.400	0.65 (+62.5%)
@ 20 photos	0.552	0.73 (+31.5%)	0.348	0.58 (+67.2%)
@ 30 photos	0.506	0.67 (+32.7%)	0.341	0.53 (+55.2%)

5 Conclusion

In this paper, we have proposed a structured document formalism of temporal events for organizing and representing home photos. We defined how to create the temporal events from a collection of temporally tagged home photos, and we described several ways to create the index corresponding to these temporal events. We formalized and developed a new way to retrieve photos using both image content and temporal context (the temporal events) and demonstrated its effectiveness and practicality for photo retrieval on a collection of 2400 home photos using 26 queries by example.

The results obtained are very promising, and we will consider in future works the retrieval of photos belonged to an entire event. Other future directions will focus on other approaches to find one or more representative image of events, in a way to avoid the use of all the images to represent the index of each temporal event.

References

[1] S.-F. Chang and J. Smith, Searching for Images and Videos on the World-Wide Web, *Technical Report #459-96-25*, Columbia Univ., Aug. 1996.

[2] Y. Chiaramella and P. Mulhem and F. Fourel, A Model for Multimedia Information Retrieval, Technical Report 4-96, *FERMI ESPRIT BRA 8134*, 1996.

[3] Cox, M. Miller, T. Minka, T. Papathomas, and P. N. Yianilos, The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments, *IEEE Tran. on Image Processing* **9**(1): 20-37, 2000.

[4] N. Craswell, D. Hawking and S. Robertson, Effective site finding using link anchor information, *ACM SIGIR*, New Orleans, 2001.

[5] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[6] M. Dunlop, *Multimedia Information Retrieval*, PhD thesis, Department of Computer Science, University of Glasgow, 1991.

[7] A. Graham, H. Garcia-Molina, A. Paepcke and T. Winograd, Time as an Essence for Photo Browsing Through Personal Digital Libraries, *ACM JCDL*, USA, pp.326-335, 2002.

[8] M. Lalmas, Dempster-Shafer's Theory of Evidence Applied to Structured Documents Modelling Uncertainty, *ACM SIGIR*, USA, pp.110-118, 1997.

[9] M. Marchiori, The Quest for Correct Information on the Web: Hyper Search Engines, *6th Intl. World Wide Web Conference*, California, U.S.A., pp.265-276, 1997.

[10] P. Mulhem and J.H. Lim, Symbolic photo content-based retrieval, *ACM CIKM*, McLean, VA, USA, Nov. 4-9, pp. 94-101, 2002.

[11] S. H. Myaeng, D.-H. Juang, M.-S. Kim and Z.-C. Zhoo, A Flexible Model for Retrieval of SGML Documents, *ACM SIGIR*, Australia, pp. 138-145, 1998.

[12] J. Platt, AutoAlbum: Clustering Digital Photographs Using Probabilistic Model Merging, *IEEE Workshop on Content-Based Access of Image and Video Libraries 2000*, pp. 96-100, 2000.

[13] J. C. Platt, M. Czerwinski, B. Field, PhotoTOC: Automatic Clustering for Browsing Personal Photographs, Microsoft Research Technical Report MSR-TR-2002-17, (2002).

[14] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, Cubic Spline Interpolation, Sub-Chapter 3.3 of Numerical Recipes in C – The Art of Scientific Computing, Second Edition, Cambridge Universoty Press, 1993, pp. 113-117.

[15] K. Rodden, How do people organise their photograpghs? *BCS IRSG 21st Annual Colloquium on Information Retrieval Research*, Glasgow, April 1999.

[16] K. Rodden and K. Wood, How People Manage Their Digital Photos? *ACM CHI*, USA, 2003 (to appear).

[17] J. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley Publisher, 1984.

[18] Y. Sun, H. Zhang, L. Zhang and M. Li, MyPhotos – A System for Home Management and Processing, *ACM Multimedia*, France, pp. 81-83, 2002.

[19] R. Wilkinson, Effective Retrieval of Structured Documents, *ACM SIGIR* 1994, Ireland, pp.311-327, 1994.