Jin L, Liu GL, Zhao L *et al.* PhotoPrev: Unifying context and content cues to enhance personal photo revisitation. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 30(3): 453–466 May 2015. DOI 10.1007/s11390-015-1536-z

PhotoPrev: Unifying Context and Content Cues to Enhance Personal Photo Revisitation

Li Jin (金 力), Gang-Li Liu (刘钢利), Liang Zhao (赵 靓), and Ling Feng (冯 铃), *Senior Member, IEEE*

Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology Tsinghua University, Beijing 100084, China

E-mail: {l-jin12, gl-liu13}@mails.tsinghua.edu.cn; zhaoliang0415@gmail.com; fengling@tsinghua.edu.cn

Received December 1, 2014; revised March 16, 2015.

Abstract Personal photo revisitation on smart phones is a common yet uneasy task for users due to the large volume of photos taken in daily life. Inspired by the human memory and its natural recall characteristics, we build a personal photo revisitation tool, PhotoPrev, to facilitate users to revisit previous photos through associated memory cues. To mimic users' episodic memory recall, we present a way to automatically generate an abundance of related contextual metadata (e.g., weather, temperature) and organize them as context lattices for each photo in a life cycle. Meanwhile, photo content (e.g., object, text) is extracted and managed in a weighted term list, which corresponds to semantic memory. A threshold algorithm based photo revisitation framework for context- and content-based keyword search on a personal photo collection, together with a user feedback mechanism, is also given. We evaluate the scalability on a large synthetic dataset by crawling users' photos from Flickr, and a 12-week user study demonstrates the feasibility and effectiveness of our photo revisitation strategies.

Keywords personal photo revisitation, memory cue, user feedback

1 Introduction

Nowadays, as the rate of digital acquisition rises, the capacity of storage becomes larger, and taking photos gets easier, we are inching closer to Vannevar Bush's 1945 Memex vision of storing a lifetime's worth of photos. With smart phones, people can record life in a variety of contexts, such as attending a conference, being on vacation, and joining a party. Facebook reveals that we daily upload a whopping 350 million public photos to the social network[①]. Meanwhile, there will be more private photos kept in users' smart phones.

The explosion in the amount of personal digital photo collections is beyond the abilities of individuals to easily manage and understand their own photos, which has made revisiting certain targets become time-consuming and boring. Personal photo revisitation faces grand challenges especially on the narrow screen of smart phones. To illustrate, let us look at the following two real photo revisitation scenarios.

Case 1. I once found a good solution to the research topic on "salient region detection" when I attended a lecture a few months ago. At that time, I took a photo of that slide using iPhone. Now I encounter a problem and want to refer to that photo. However, it turns out to be hard to re-localize the exact photo from dozens of photos in my iPhone.

Case 2. I took a large number of photos about temples using iPhone when I was on vacation. I once took a photo about a famous temple during a trip to India. It would be handy to return that exact photo rather than a bundle of photos about temples to recognize.

Photo revisitation is different from photo finding. There is uncertainty in the latter process because users

Regular Paper

Special Section on Computational Visual Media

The work was supported by the National Natural Science Foundation of China under Grant Nos. 61373022, 61073004, and the National Basic Research 973 Program of China under Grant No. 2011CB302203-2.

①https://fbcdn-dragon-a.akamaihd.net/hphotos-ak-prn1/851575_520797877991079_393255490_n.pdf, Sept. 2014.

do not know enough information, while revisitation is a more directed process as users have already taken or browsed the photos before. A general way to support personal photo revisitation is to maintain photos' associated context and content information. However, how to manage such photos' associated information to mimic users' retrieval and recall mechanism is still a challenging research topic, which can make revisitation process more personalized and better serve users to improve satisfaction.

In this paper, we consider the problem defined as follows: given a large personal photo collection $\mathcal{P}$ taken by smart phones, PhotoPrev returns top-k photos after typing in context- and content-based keyword query with fast response time and good revisit quality. The salient character of our problem definition lies in that we aim to mimic natural retrieval and recall mechanism of episodic and semantic memory. In personal photo revisitation, the episodic memory is related to the photo's associated context, while the semantic memory is related to the photo's content.

The main contributions of this paper are as follows.

• PhotoPrev automatically acquires and represents each access context and photo content as a context lattice and a weighted term list respectively, linked to the possible to-be-revisited photo. The constructed context and content memory are able to evolve as time elapses to mimic degradation mechanism.

• PhotoPrev periodically learns from the feedbacks of a user's keyword-based query search, and adapts to revisit habit accordingly. The feedback adaptation adjusts parameters during memory management to offer personalized memory retrieval.

• We report the findings of a 12-week user study with our prototype on personal photo revisitation.

The rest of the paper is organized as follows. We briefly review some closely related work in Section 2. We address the construction and management of associated context and content in Section 3. We present a threshold algorithm based photo revisitation framework in Section 4, and introduce user feedback adaptation in Section 5. Section 6 describes the design and implementation of PhotoPrev, whose performance is evaluated in Section 7. We finally discuss the limitations of our prototype in Section 8, and conclude the paper in Section 9.

2 Related Work

2.1 Context-Based Photo Revisitation

There is a large amount of work on context-based photo revisitation, which mainly explored how to generate an abundance of associated contextual metadata and manage them using effective methods. PhotoCompas[1] proposes browseable location and activity hierarchies to organize each personal photo collection. Naaman *et al.*[2] extended PhotoCompas to add more context (e.g., light status), and conducted a detailed user study to demonstrate which categories of contextual metadata are most useful when revisiting photos. Bearing similarities to the previous work[2], Cao *et al.*[3] proposed multi-level annotation hierarchy considering more semantic information. Joshi and Luo[4] proposed a classification algorithm to infer generic activities through combining visual and geo-tag information. PhotoMap[5] provides an automatic spatio-temporal annotation for mobile photos, which combines web services and social network profiles to build context ontology. Viana *et al.*[6] regarded photo context as a bag of words to realize keyword-based retrieval process by extending the traditional vector space model.

2.2 Content-Based Photo Revisitation

Object Recognition. For object recognition, the research work can be mainly divided into two types of approaches: parametric approaches that consist of learning generative/discriminative models, and nonparametric approaches that rely on image retrieval and matching. Among parametric approaches, Crandall *et al.*[7] proposed a class of statistical models for part-based object recognition based on the degree of spatial structure. Dalal and Triggs[8] studied proper feature parameters for robust visual object recognition with histograms of oriented gradient (HOG) descriptors. In addition, Felzenszwalb *et al.*[9-10] designed similar constellation models to regard objects as ensembles of parts. These methods focus on articulated objects, which are mostly rigid and susceptible to little or no deformation. Hu *et al.*[11] proposed an unsupervised feature selection approach based on Bag-of-Visual-Words model for high-dimensional object indexing. Zhou *et al.*[12] proposed a spatial context coding strategy for visual matching verification, which could decrease the false local matches between images. Shotton *et al.*[13] proposed a discriminative model that combines texture-layout filters with

lower-level image features to realize multi-class object recognition and segmentation. Among nonparametric approaches, Hu *et al.*[14] detailed the recent research about the object extraction and matching to assist visual media analysis. Frome *et al.*[15] proposed local perceptual distance functions to compute the distance between a query image and images in the training set, which subsequently cast votes to infer the object class of the query. Russell *et al.*[16] built a probabilistic model to transfer the labels from a densely labeled image database (e.g., LabelMe[17]) to the input image based on the nearest neighbor. Liu *et al.*[18] extended the previous work[16] and proposed a method to improve recognition accuracy, which involves firstly a retrieval step on a large database of annotated images using a modified version of SIFT flow[19]. Then a Markov random field framework is applied to integrate multiple cues to segment and recognize the query image. Cao *et al.*[20] proposed a geometric method between images for similarity judgment in high-dimensional space.

Text Recognition. Numerous research work deals with text recognition from photos, which can be broadly categorized into two groups: texture-based methods and region-based methods. Through scanning image at various scales, texture-based methods extract a number of text properties, such as the distribution of wavelet coefficients[21], high variance of intensity, low gradients above and below text[22-23]. Region-based methods detect connected components, which group pixels with certain properties, such as approximately constant color, and stoke width[24-25]. Matas *et al.*[26] proposed maximally stable extremal regions (MSERs), which are particular cases of extremal regions (ERs) whose size remains virtually unchanged over a range of thresholds. However, MSERs still have problems on blurry images or characters with low contrast. Neumann and Matas[27] dropped the stability requirement of MSERs and proposed a classification method to optimally select class-specific (not necessarily stable) ERs to enhance the robustness. To overcome the effect of affine or deformation on text extraction, Zhang *et al.*[28] proposed an efficient algorithm to detect and rectify texts in arbitrary orientations against complex background.

For user interaction, Sketch2Photo system[29] generates photo-realistic pictures from the user's sketch of a scene with text label annotated objects. Candidate images matching the sketch and text labels can be obtained by searching the Internet. Then a hybrid image blending algorithm is presented to realize seamless image composition. ShadowDraw[30] guides the freeform drawing of objects for users by providing shadows derived from images in real time. In this paper, we adopt keyword-based traditional search considering mobile phone's narrow screen for personal photo collections.

3 PhotoPrev Backend

To prepare photo revisitation via captured context and photo content, PhotoPrev acquires and manages associated context as well as photo content information to mimic human memory upon a photo access.

3.1 Context Memory

Given time and location information about digital photos, we can automatically generate an abundance of related contextual metadata using web services to assist personalized photo revisitation. In this subsection, we perform two tasks, which are the acquisition of associated context cues and the dynamic management of context memory.

3.1.1 Context Cues Acquisition Module

When a user takes photos by a smart phone, time and location can be automatically recorded. Access time c_{time} is determinate. Access location c_{loc} is obtained based on the IP address or possible GPS information of the smart phone if available. We infer the user's generic activities c_{act}s by leveraging the inherent patterns of association based on corresponding geo-tags (e.g., POIs) and visual concepts by employing state-of-the-art visual detection algorithms[18,27]. Firstly, activity classes' descriptions and visual concepts associated with them are defined based on the latent similarity proposed by [4], where a practical criterion is based on the tags' popularity in Flickr②. Then, a classification algorithm[4] combining visual and geo-tag information is adopted to label activity class on each photo with high association probability. In addition, the time and the location where photo p_i was taken allow us to retrieve archival data from weather stations③ which are local to p_i's exposure location. Similarly, we automatically obtain other useful contextual metadata categories (e.g., light status, time zone, temperature) based on how well they were remembered[2] as context cues for human memory.

②http://www.flickr.com, Sept. 2014.

③http://cdc.nmic.cn/home.do, Sept. 2014.

Fig.1. Example of object recognition and text localization for a personal photo collection.

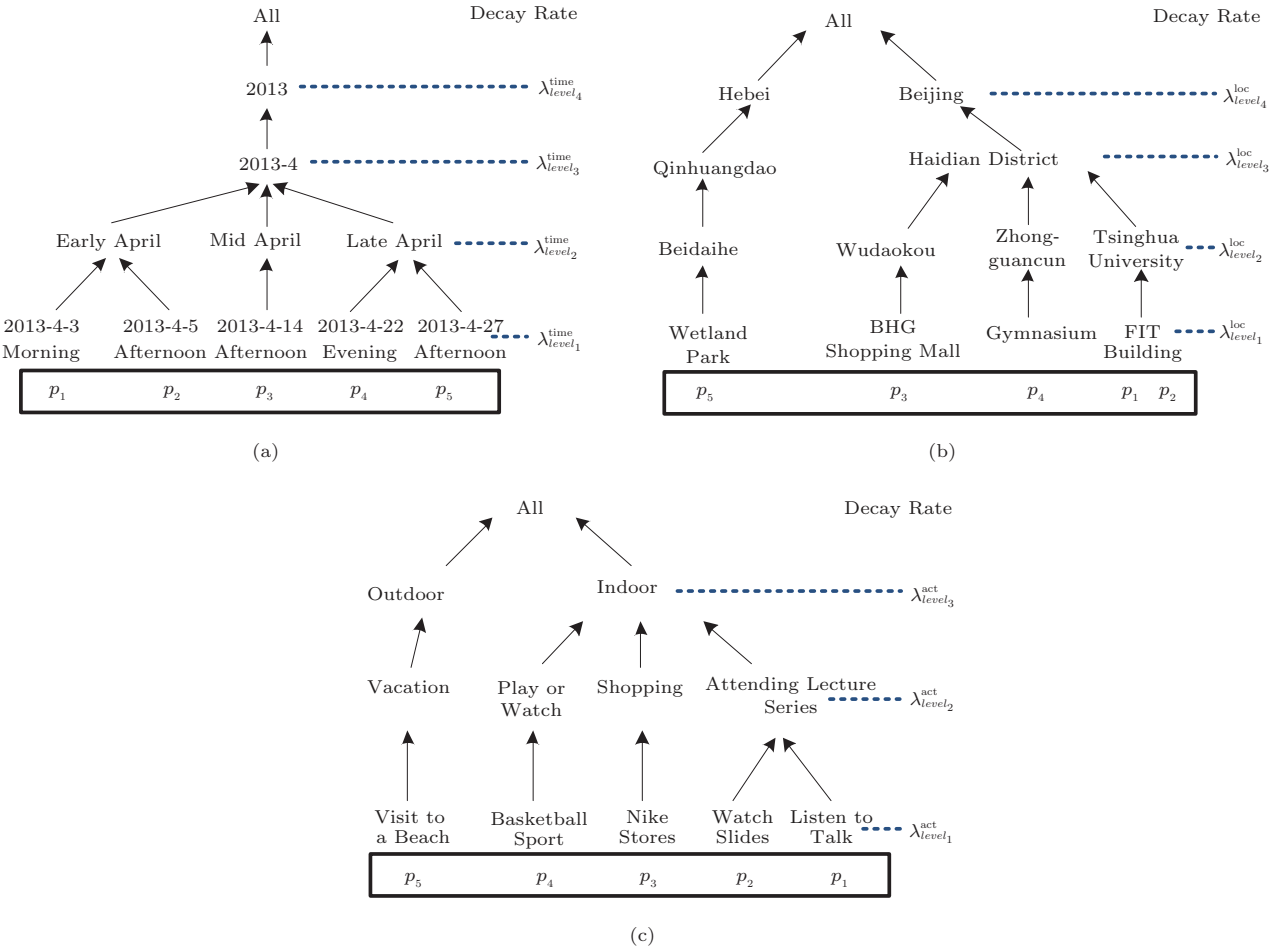


Fig.2. Contextual hierarchies of (a) time, (b) location, and (c) activity for a personal photo collection.

Example 1. Considering photo p_5 in Fig.1, it was taken in "Wetland Park" and contains person, sea, sky and grass. Then we can infer that the activity class label corresponds to "visit to a beach" and "on vacation", as shown in Fig.2(c).

Definition 1 (Context Lattice). *Access context $\mathcal{A}$ of each photo is comprised of n contextual attributes $(\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n)$, where each attribute is segmented into hierarchies using corresponding concepts from popular knowledge base Yago. The hierarchy of context attribute $\mathcal{A}_i$ can be viewed as a lattice $(Dom(\mathcal{A}_i), H, \prec_h)$, where $H = (h_1, h_2, \ldots, h_m)$ of m levels corresponds to $level_{\mathrm{Id}}$ $(1, 2, \ldots, m)$, and $\prec_h$ is a partial order among the levels of H. $level_{\mathrm{Id}} \in \{1, 2, ..., m\}$ is the number of*

each hierarchical level, and $level_{\mathrm{Id}}$ *of the bottom level is* 1*. Assume context values* v' *and* v *are at different hierarchical levels,* v' *is called an* ancestor *of* v*, denoted as* $v \prec_a v'$*, only if there exists an upward path from* v *to* v'*.*

Example 2. Considering location context in Fig.2(b), we build a 4-leveled abstraction hierarchy H, where "Wetland Park"$\prec_a$ "Beidaihe"$\prec_a$ "Qinhuangdao"$\prec_a$"Hebei".

Therefore, a context instance of photo p is an instantiation of its n contextual attributes, represented as a multi-dimensional vector $\mathcal{C} = (c_1, c_2, \ldots, c_n)$, where c_i is the *i-th* context lattice for corresponding contextual attribute.

3.1.2 Context Memory Management Module

To mimic the characteristic of human brain memory that the majority of personal photos' associated context instances will gradually degrade and disappear in the end, we propose a dynamic life-cycle decay policy based on cognitive psychology studies[31-34] for context hierarchies of a personal photo collection as shown in Fig.2. When a context instance has all its attribute values decayed to root node All, we think the context instance has been forgotten.

Definition 2 (Retention Strength). *Let* $\mathcal{A}_i$ *be a contextual attribute with a value* $v \in Dom(\mathcal{A}_i)$*. The retention strength of* v*, denoted as* $R(\mathcal{A}_i, v, t)$*, is a real number* $R \in [0, 1]$*, characterizing the memorized state of* v *as a function of the exponential in the square root of elapsing time* $t - T_{n-1}^{\mathcal{A}_i}$ *(also called age)*[32]*:*

$$R(\mathcal{A}_i, v, t) = r_0 \times e^{-\lambda_{level_n}^{\mathcal{A}_i}\sqrt{t-T_{n-1}^{\mathcal{A}_i}}}, \quad if \quad t > T_{n-1}^{\mathcal{A}_i},$$

where r_0 *is the initial value of retention strength,* $\lambda_{level_n}^{\mathcal{A}_i} = \frac{1}{T_n^{\mathcal{A}_i} - T_{n-1}^{\mathcal{A}_i}}$[35] *is the decay rate at* $level_n$ *in context hierarchy of* $\mathcal{A}_i$ *(e.g., location),* $T_{n-1}^{\mathcal{A}_i}$ *is the initial day when the retention strength of context node* v *in* $level_n$ *begins to decay. If elapsing time* t *is more than* $T_n^{\mathcal{A}_i}$*, the retention strength of* v *in* $level_{n+1}$ *begins to decay along the hierarchical path.*

Note that decay rate $\lambda_{level_n}^{\mathcal{A}_i}$ is user-dependent, which is firstly assigned to an initial value, and then adjusted based on user feedback. The bigger decay rate $\lambda_{level_n}^{\mathcal{A}_i}$ is, the more memory retention strength $R(\mathcal{A}_i, v, t)$ drops, signifying the fast context value v degrades. With time elapsed, users can only remember some general context values of previous accessed photos. Therefore, the hierarchies of *time, location,* and *activity* in the context memory evolve dynamically in life cycles to reflect the gradual degradation of human's context memorization as well as the generalized context-based keyword queries that users will use for recall[31], as shown in Fig.2.

Example 3. Consider a context value v = "Wetland Park" of location hierarchy in Fig.2. Assume the initial retention strength $r_0 = 1.0$, $T_0^{\mathrm{loc}} = 0$ and $T_1^{\mathrm{loc}} = 15$. After 10 days ($t = 10$), v's retention strength will become $R(\mathcal{A}_{\mathrm{loc}}, v, t) = 1.0 \times e^{-\frac{1}{T_1^{\mathrm{loc}} - T_0^{\mathrm{loc}}}\sqrt{t-T_0^{\mathrm{loc}}}} = 1.0 \times e^{-\frac{1}{15}\sqrt{10}} = 0.809\,9$, where $t - T_0^{\mathrm{loc}}$ is to calculate the elapsing time of context node v in $level_1$.

We conduct a user study to determine the initial value of decay parameters for different context attributes in Subsection 7.2.1. And user feedback adaptation to adjust decay parameters for personalized revisitation is described in Section 5. To organize the context lattices of personal photo collection, we adopt Dewey code and build inverted index for multi-dimensional context vectors $\mathcal{C}_{\mathrm{set}}$ to facilitate context-based keyword query $Q.context$.

3.2 Content Memory

When a user takes/browses a photo, he/she may focus on some interesting parts, i.e., object and text, which leave a deep impression. Therefore, except for associated context, we should analyze and capture useful content cues to construct content memory for personal photo revisitation.

3.2.1 Content Cues Extraction Module

Label transfer[18] can achieve a good performance on object recognition; however, text in photos can also be regarded as important content cues. Therefore, we firstly use photo OCR technology④ to localize and recognize text, and then adopt label transfer to realize nonparametric scene parsing based on open source code package⑤. Stop words and words not in WordNet⑥ for photo text are removed as shown in Fig.3(a).

For photo OCR technology, the pipeline can be divided into three steps: text localization, character segmentation and character recognition. For text localization, it can be regarded as how to efficiently select a

④https://code.google.com/p/tesseract-ocr/, Sept. 2014.

⑤http://people.csail.mit.edu/celiu/LabelTransfer/, Sept. 2014.

⑥http://wordnet.princeton.edu/, Sept. 2014.

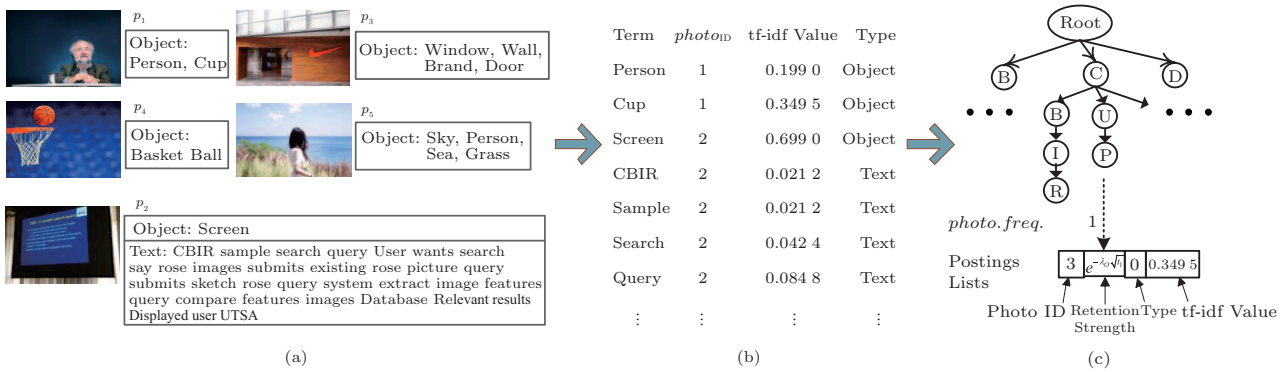


Fig.3. A weighted term list example for a personal photo collection and index. (a) Photo content extraction. (b) Weighted term list. (c) Trie tree with inverted index.

set of extremal regions (ERs), where an ER r is a region whose outer boundary pixels have strictly higher values than the region itself[27]. A two-stage sequential classifier is used to calculate ERs by combining a lot of features. In the first stage, a real AdaBoost classifier with decision trees is used with the features: aspect ratio, compactness, number of holes and a horizontal crossings feature. In the second stage, an SVM classifier with the RBF kernel is used to classify ERs into character and non-character classes considering more informative and more computationally expensive features: hole area ratio, convex hull ratio, and the number of outer boundary inflexion points. Then we group ERs into words and select the most probable character segmentation. Finally, text can be recognized in an OCR training stage.

To realize scene parsing for a photo p_i, label transfer matches p_i's visual objects to the images in a database (e.g., LabelMe[17]). If images in such databases are annotated with object category labels that are semantically meaningful, it will transfer labels of images in the database to parse the input. SIFT flow[19] is adopted to establish semantically meaningful correspondences between two images by matching local SIFT descriptors. A coarse-to-fine pyramid SIFT flow matching algorithm is proposed to estimate the flow at a coarse level of image grid, and then gradually propagate and refine the flow from coarse to fine[19]. Finally a probabilistic Markov random field model is adopted to integrate multiple labels, the prior information of object category, and the spatial smoothness of the annotation to parse p_i.

Our object detection method is a large database-driven approach, whose unique characteristic is openness. When adding more images of the new categories into database, it does not require additional training. Meanwhile, label transfer can predict the right object categories in the input image with a segmentation fit to image boundary, even though the best match may look different from the input. An example of scene parsing for a set of personal photos using label transfer is illustrated in Fig.1.

3.2.2 *Content Memory Management Module*

Extracted terms evolve dynamically in life cycles to reflect the gradual degradation of human's content memory, which users can input content-based keyword query $Q.content$ to revisit. And the retention strength of each term will progressively decay with time at $R(\mathcal{T}_\zeta, c_{\text{term}}, t) = r_0 \times e^{-\lambda_\zeta \sqrt{t}}$, where $\lambda_\zeta = \frac{1}{T_\zeta}$ is the initial decay rate, and $\zeta \in \{object, text\}$.

For recognized object and text, we treat each photo as two document types, i.e., Doc_{obj} and Doc_{text}. Then we can calculate the tf-idf value to measure the memory influence. For a term c_{term}, its tf-idf value is $tfidf(c_{\text{term}}) = \log \frac{n}{df(c_{\text{term}})} \times tf(c_{\text{term}})$, where $df(c_{\text{term}})$ is the number of accessed photos containing c_{term}, n is the total number of accessed photos, and $tf(c_{\text{term}})$ is the occurring number of c_{term} in the current accessed photos. Note that we just consider the relevant terms in the same document type of photos when calculating c_{term}'s tf-idf value.

To gain the speed benefits of indexing at retrieval time, we apply Trie tree to organize the extracted term lists $\mathcal{I}$ based on the longest common prefix. For each tree node, inverted index is built to store a mapping from extracted term lists in advance. Within a to-be-revisited personal photo collection $\mathcal{P}$, we assume that each photo has a unique serial number, known as photo identifier ($photo_{\text{ID}}$). During index construction, the input is $\mathcal{I}$ for $\mathcal{P}$, and then we insert the terms into the Trie tree. Meanwhile, instances of the same term are

grouped together, and the result is split into a dictionary and postings as shown in Fig.3(c). The dictionary records some statistics, such as the number of photos that contain each term ($photo.freq.$), which also corresponds to the length of each postings list ρ_{list}. And ρ_{list} stores a list of pairs of photo identifier $photo_{\text{ID}}$, retention strength R, document type Doc_ζ and tf-idf value $tfidf(c_{\text{term}})$ for each extracted term.

4 Keyword-Based Query Search Module

A keyword-based query for photo revisitation can be denoted as a function $RF(Q, CM)$, where $Q = \{\mathcal{K}_Q, k\}$ is the query request containing a set of context-based keywords $Q.context$ and content-based keywords $Q.content$, and CM is the query target that is the memory snapshot, which dynamically evolves in life cycles according to query user's memorization strength. The result of Q upon CM is the top-k ranked photos $(photo_1, photo_2, \ldots, photo_k)$, whose ranking score is determined by a context- and content-based similarity function between Q and CM.

Definition 3 (Context- and Content-Based Similarity). *Given a user query Q and human memory snapshot CM of photo p, the context- and content-based similarity between Q and p is defined as*

$$Sim(Q,p) = \alpha Sim_{\mathcal{A}}(Q, p.CM_{\mathcal{A}}) + (1-\alpha) Sim_{\mathcal{T}}(Q, p.CM_{\mathcal{T}}),$$

where $Sim_{\mathcal{A}}(Q, p.CM_{\mathcal{A}})$ is context-based similarity and $Sim_{\mathcal{T}}(Q, p.CM_{\mathcal{T}})$ is content-based similarity. A parameter α is to balance relative importance between the associated context and content cues.

Definition 4 (Context-Based Similarity). *Given context-based keywords $Q.context = \{\mathcal{K}_{\mathcal{A}}\}$ and context memory snapshot $CM_{\mathcal{A}}$ of photo p, the context-based similarity between Q and $p.CM_{\mathcal{A}}$ is defined as:*

$$Sim_{\mathcal{A}}(Q, p.CM_{\mathcal{A}}) = \sqrt{\frac{1}{|\mathcal{K}_{\mathcal{A}}|} \sum_{i=1}^{|\mathcal{K}_{\mathcal{A}}|} (R^2(\mathcal{A}, q_i, t))},$$

where q_i is the i-th context-based keyword of $\mathcal{K}_{\mathcal{A}}$.

Definition 5 (Content-Based Similarity). *Given content-based keywords $Q.content = \{\mathcal{K}_{\mathcal{T}}\}$ and content memory snapshot $CM_{\mathcal{T}}$ of photo p, the content-based similarity between Q and $p.CM_{\mathcal{T}}$ is defined as:*

$$Sim_{\mathcal{T}}(Q, p.CM_{\mathcal{T}}) = \sqrt{\frac{1}{|\mathcal{K}_{\mathcal{T}}|} \sum_{i=1}^{|\mathcal{K}_{\mathcal{T}}|} (tfidf(q_i) \times R(\tau, q_i, t))^2},$$

where q_i is the i-th content-based keyword of $\mathcal{K}_{\mathcal{T}}$.

Inspired by the threshold algorithm (TA) algorithm[36], we propose a TA-based framework to efficiently find similar photos for a query. The basic idea is that, if photo p is a top-k result of query Q, then either its context or content cues should be similar enough. Thus, by building inverted index for context and content memory separately, we can quickly find photos with large context- and content-based similarity. We take these photos as candidates and then verify them to generate the final results. As shown in Algorithm 1, we use Υ to dynamically keep k objects with the current highest similarity. θ_Q is the lowest value in Q (Line 2). At each loop, we incrementally find photo $p_{\mathcal{A}}$ with the current highest context-based similarity (line 6). If its context- and content-based similarity $Sim(Q, p_{\mathcal{A}.CM})$ is larger than θ_Q, we add $p_{\mathcal{A}}$ to Υ and update θ_Q (lines 7∼9). Similarly, we incrementally find photo $p_{\mathcal{T}}$ with the current highest content-based similarity and update θ_Q (lines 10∼13). A threshold θ_{TA} is maintained to indicate the maximum similarity for unvisited photos and updated at the end of each loop (lines 14∼15). If $\theta_Q > \theta_{\text{TA}}$, we can return the k photos in Υ as results because none of the unvisited photos may get a higher similarity than θ_Q.

Algorithm 1. TA-Based Top-k Photo Revisitation
Input: a revisit request Q formalized as $\{\mathcal{K}_Q, k, \alpha\}$; human memory snapshot for context and content cues $CM = \{CM_{\mathcal{A}}, CM_{\mathcal{T}}\}$ of photo collection $\mathcal{P}$
Output: Υ: k most similar photos

```
1  begin
2      Υ ← ∅, θ_TA ← 1, θ_Q ← 0;
3      while true do
4          if θ_Q > θ_TA then
5              Return Υ;
6          photo p_A ← BestContextSearch(K_Q, P_.CM_A);
7          if |Q| < k or Sim(Q, p_A.CM) > θ_Q then
8              Add photo p_A to Υ;
9              Update threshold θ_Q;
10         photo p_τ ← BestContentSearch(K_Q, P_.CM_τ);
11         if |Q| < k or Sim(Q, p_T.CM) > θ_Q then
12             Add photo p_τ to Υ;
13             Update threshold θ_Q;
14         θ_TA ← αSim_A(Q, p_A.CM) + (1 − α)×
15                Sim_τ(Q, p_T.CM);
```

5 User Feedback Adaptation Module

As the outcome of context and content memory management will directly impact the actions of a user's

photo revisitation by keyword-based query Q, the revisit feedback should be taken into account in the ongoing memory management to make the process more personalized.

Adjustments. To adjust the decay rate $\lambda^{\mathcal{A}_i}$ of the i-th associated context attribute $\mathcal{A}_i$, we need to count the average days $T^{\mathcal{A}_i}$. After that, a user's context-based keywords begin to become general along the hierarchical path. Note that $Q.context$ may contain context-based keywords from different levels of attribute set $\mathcal{A}$, and each context attribute has different decay rates at different levels. The parameter $D_{ays}(Q.context, \mathcal{A}_i, n)$ is the days between browsing a web page and revisiting it by matching context node in $level_n$ for $\mathcal{A}_i$. Then we can assume users' revisit habit satisfies a normal distribution $D_{ays} \sim \mathcal{N}^{\mathcal{A}}(\mu, \delta)$. And $T_n^{\mathcal{A}_i} = \mu_n^{\mathcal{A}_i} + 2\delta_n^{\mathcal{A}_i}$ is calculated by using the upper bound of $\mathcal{N}^{\mathcal{A}_i}$. For example, considering location context $\mathcal{A}_{\text{loc}}$ has four levels in Fig.2 and $Q.context$ = "april tsinghua" involves context-based keyword "tsinghua" from level 2 of $\mathcal{A}_{\text{loc}}$ in Fig.4(b), we record the days $D_{ays}(Q.context, \mathcal{A}_{\text{loc}}, 1)$ between browsing/taking a photo and revisiting it. After calculating the statistical distribution $\mathcal{N}^{\mathcal{A}_{\text{loc}}}$ and estimating upper bound $T_1^{\mathcal{A}_{\text{loc}}}$, we can update the $\lambda_1^{\mathcal{A}_{\text{loc}}} = \frac{1}{T_1^{\mathcal{A}_{\text{loc}}} - T_0^{\mathcal{A}_{\text{loc}}}}$ of level 1. Then we can determine T_2^{loc} and update λ_2^{loc} based on matching nodes against $Q.context$ in a similar manner. To adjust decay rate λ_ζ of associated content terms, we take the top $\tau_{\mathcal{E}}$ of terms as a set $\mathcal{E}$ based on tf-idf value. After counting the average days $D_{ays}(Q.content, \mathcal{E})$ when a user's content-based keywords $Q.content$ do not belong to $\mathcal{E}$, we estimate the upper bound T_ζ of corresponding statistical distribution $\mathcal{N}^\zeta$ and then update $\lambda_\zeta = \frac{1}{T_\zeta}$. The adjustment of decay rate is to catch the user's memory behavior and realize proper memory matching.

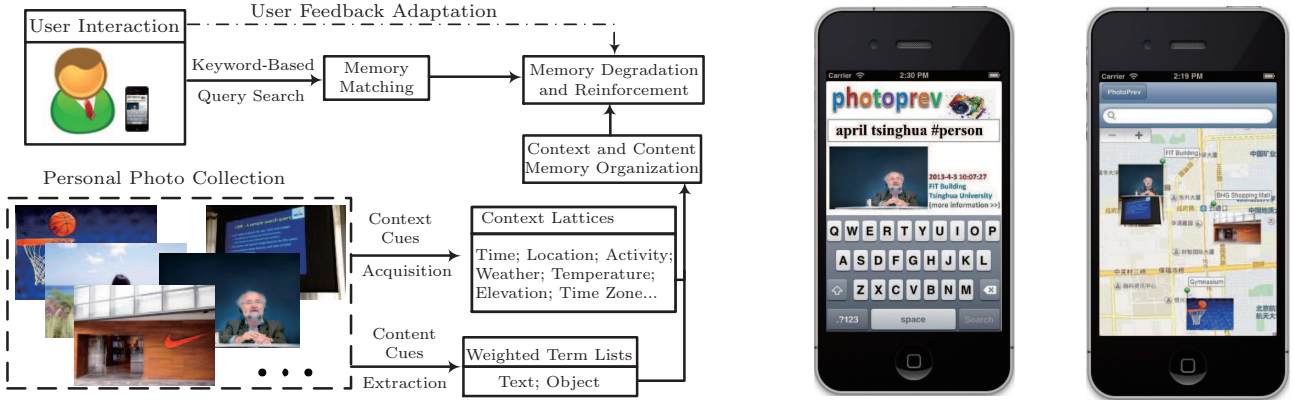
Reinforcement. Because recall actions can often refresh users' memory, during evolution process, certain parts of context and content memory are reinforced due to users' revisit actions. For context memory, if a user types in a context value in the context lattices, its possibly degraded retention strength is reset to the original one. The decay starting time for its located level is meanwhile reset to the current time. For content memory, we update the matching terms' retention strength in a similar fashion.

6 User Interaction

Users interact with PhotoPrev during their photo access phase and photo revisitation phase.

When a user takes a photo by a smart phone, PhotoPrev will automatically perform both context cues acquisition and content cues extraction, and then manage them into context lattices and weighted term lists.

During the photo revisitation phase, PhotoPrev provides two types of search interfaces for a user to select, as shown in Fig.4(b). For example, the user can type in the following context-based keywords "april tsinghua", and content-based keyword "person". Here, the user's context input may not be as precise as the original context due to the natural fading of human memory as time goes by. In the above case, instead of the exact time and location context "2014-4-3 morning FIT Building", the user may only remember that this talk happened in Tsinghua University during April. Through build-



(a) (b)

Fig.4. PhotoPrev: unifying context and content cues to enhance personal photo revisitation. (a) PhotoPrev architecture. (b) Mobile UI.

ing context hierarchies, PhotoPrev can identify those closely matching context units from personal context memory and present an efficient algorithm for personal photo revisitation. When the user wants to input a query, he/she can add "#" to help PhotoPrev classify context- and content-based keywords. Otherwise, PhotoPrev will automatically label each keyword based on CRF model[37] and recommend appropriate queries for the user to confirm.

After entering the revisit request and clicking the "search" button, PhotoPrev seeks and returns a ranked list of top-k photos in Fig.4(b). The user can double-click a returned photo to see detailed information and then confirm it. To protect user privacy, the user can also mark some photos which will not be submitted to PhotoPrev.

7 Evaluation

Two sets of experiments are performed to examine the performance of PhotoPrev. The first experiment aims to study its scalability issue on a large synthetic dataset, and the second one aims at its applicability and acceptance issues through a 12-week real user study. Two performance measurements (revisit response time and revisit quality) are adopted throughout the experiments. Revisit response time is used to test the system average response time when users input queries on a large context and content memory. Revisit quality is based on average revisit precision, recall, and ranking position.

7.1 Experiment on Synthetic Data

7.1.1 Synthetic Data Generation

We firstly build two extra components: 1) data simulator, to simulate the generation of personal photo collections; and 2) user simulator, to simulate the user's memory over the generated data and revisit actions, acts as a "real user". Synthetic data generation lies in the following two aspects.

Generation of Context and Content Memory. The data simulator crawls users' photos with contextual metadata (e.g., time, location) to form a dataset using Flickr API[7] from social network. Considering personal photo collection, data simulator mainly selects users who share more than 500 photos on Flickr. Then the data simulator generates $photo_{\text{num}}$ (1 k, ..., 100 k) photos, which correspond to context lattices and content term lists to mimic memory snapshot.

Generation of Revisit Requests. Every period (seven days), the user simulator formulates 10 revisit requests against above generated context and content memory. Each revisit request contains $keyword_{\text{num}}$ (2, ..., 10) keywords, which are randomly selected from the corresponding context lattices and content term lists. PhotoPrev processes the revisit requests from the user simulator periodically, and then relevant parameters are updated based on user feedback adaptation.

7.1.2 Experimental Results of Synthetic Data

In this subsection, we mainly compare the average response time of 6-month revisit requests under different parameters generated by the user simulator. The experiment is implemented in Object-C, running on iPhone 5S with iOS 7.1.

Evaluation on k. To evaluate the effect of parameter k, we fix α to 0.5, $keyword_{\text{num}}$ to 6, and vary k from 2 to 20. The result is shown in Fig.5(a), and we can see that the average response time keeps linear increasing with the increase of k. PhotoPrev scales well as $photo_{\text{num}}$ increases. The average response time with $photo_{\text{num}} = 100$ k is about 1.1 seconds, which is 2.48 times of that with $photo_{\text{num}} = 10$ k.

Evaluation on α. To evaluate the effect of parameter α, we fix k to 10, $keyword_{\text{num}}$ to 6, and vary α from 0.1 to 0.9. The result is shown in Fig.5(b), and we can see that PhotoPrev performs well when α belongs to the range of 0.6 to 0.7. It illustrates that context-based similarity accounts for larger weight than content-based similarity. Because context attributes are more plentiful and context hierarchies can do a lot of pruning operations.

Evaluation on $keyword_{\text{num}}$. To evaluate the performance under different lengths of query keywords, we fix α to 0.5, k to 10, and vary $keyword_{\text{num}}$ from 2 to 10. The result is shown in Fig.5(c), and we can see that the average response time does not increase in direct ratio along with $keyword_{\text{num}}$. When typing more context-based keywords, the number of candidate context trees will be reduced.

7.2 Experiment on User Study

7.2.1 Parameter Settings

To determine the initial decay parameters of memory cues for personal photos, we invited 44 persons (27

[7] http://www.flickr.com/groups/api, Sept. 2014.

males and 17 females, aged between 21 and 57) to conduct a user study, who always took photos by smart phones and saved more than 600 photos. To design a questionnaire for each participant, we particularly select 50 photos, whose shooting time ranges from 10 days to more than 2 years before. Among the selected photos, 40% are marked with part of content cues, and the rest are marked with part of context cues. For photos with marked cues, participants should fill in the value of other memory cues if they remember. Then we can calculate $T_i^{\mathcal{A}}$ and T_ζ as shown in Fig.6.

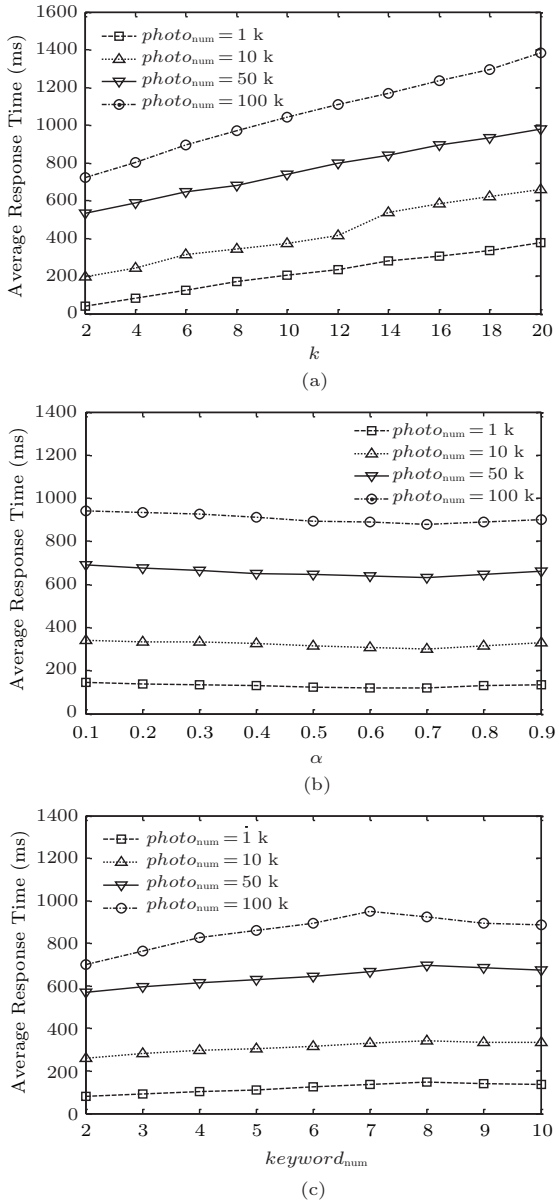


Fig.5. Experimental results on synthetic data with different parameter values. (a) $\alpha = 0.5$, $keyword_{\text{num}} = 6$. (b) $k = 10$, $keyword_{\text{num}} = 6$. (c) $\alpha = 0.5$, $k = 10$.

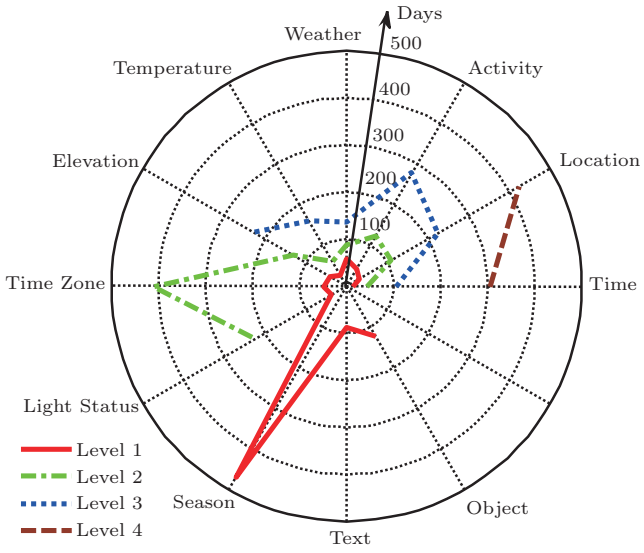


Fig.6. Statistics of $T_i^{\mathcal{A}}$ and T_ζ from questionnaire.

7.2.2 Statistics and Setup

A 12-week user study was conducted to investigate the performance of PhotoPrev in real case, with 14 participants (6 males and 8 females, aged between 21 and 41), whose iPhones were installed with PhotoPrev. During that period, participants were asked to freely revisit the personal photos using PhotoPrev, which kept the revisitation details automatically. The user study gathered 2 691 photo revisitation records in total, about 192 records per participant in average, and each participant input 16 revisit queries per week.

Considering context-based keywords $Q.context$ for user query Q, participants preferred to use location c_{loc} as context cue, which accounts for 23.2% as shown in Fig.7(a). Although text accounts for just 12.3%, we discover that the proportion of text ascends to 36.7% if to-be-revisited photos contain text. It demonstrates that photo text is also an important content cue. With time elapsing, participants are more inclined to use context-based keywords as shown in Fig.7(b), which identifies that context hierarchies are more aligned with human retrieval and recall mechanism.

7.2.3 Experimental Results of User Study

In this subsection, we mainly compare the revisit quality of PhotoPrev on a 12-week user study.

Evaluation on User Feedback Adaptation and Decay Mechanism. For feedback adaptation, PhotoPrev at first does not grasp the revisit habit of users for later revisit, and the result quality is not so good. As time

goes by, since PhotoPrev adapts to revisit habit, revisit quality gradually becomes better. Through varying $\tau_{\mathcal{E}}$, we aim to determine proper value to adapt content terms. Meanwhile, we verify decay mechanism plays a very important role in photo revisitation to improve revisit quality. For revisit quality comparison, PhotoPrev with user feedback adaptation (AdaptDecay, $\tau_{\mathcal{E}} = 20\%$) achieves (28.1%, 97.5%, 1.7) in the average precision, recall rate, ranking position compared with PhotoPrev without adaptation (DecayOnly) (24.8%, 91.8%, 3.3) and PhotoPrev without decay (NoDecay) (19.6%, 88.5%, 3.8) as shown in Fig.8.

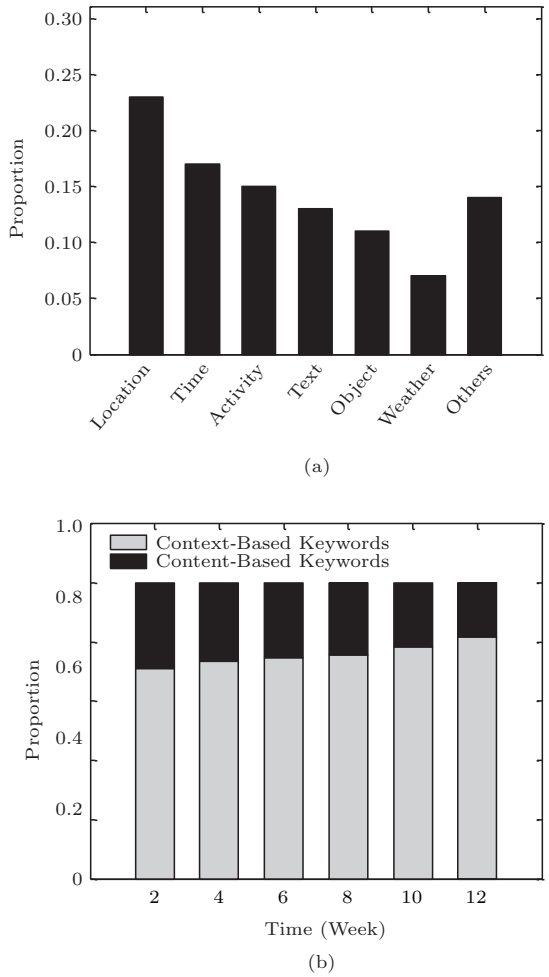


Fig.7. Statistics for user revisit request Q. (a) Proportion of using different context attributes for $Q.context$. (b) Proportion of $Q.context$ and $Q.content$ for Q.

Evaluation on Combination of Context and Content Cues. From the results presented in Fig.9, we can find that unifying context and content cues (Hybrid) to revisit delivers the best average precision, recall rate, ranking position (28.1%, 97.5%, 1.7) compared with only using context cues (ContextOnly) (18.1%, 82.9%, 3.7) and using content cues (ContentOnly) (15.6%, 76.2%, 4.6). Although PhotoPrev supports general matching for context-based keywords, and participants tended to revisit by general context-based keywords like vacation, shopping, attending lecture series and so on, the content-based keywords can narrow down the search scope and reflect the user's revisitation intention well.

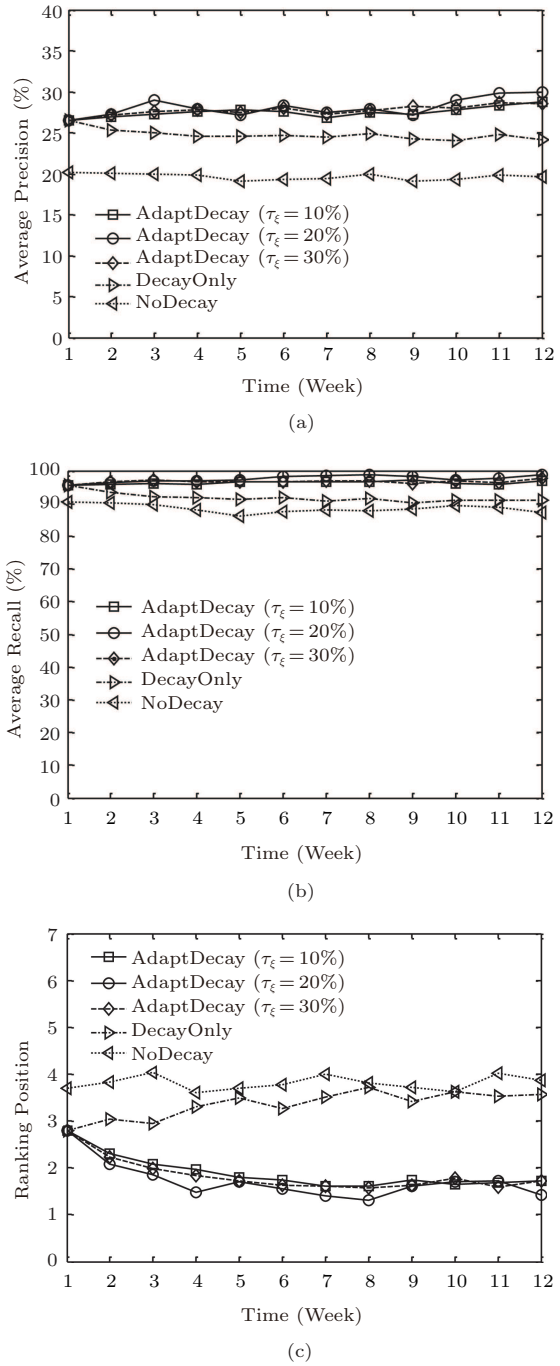


Fig.8. Comparison results with/without user feedback adaptation and decay mechanism on a 12-week user study ($k = 10$, "Hybrid").

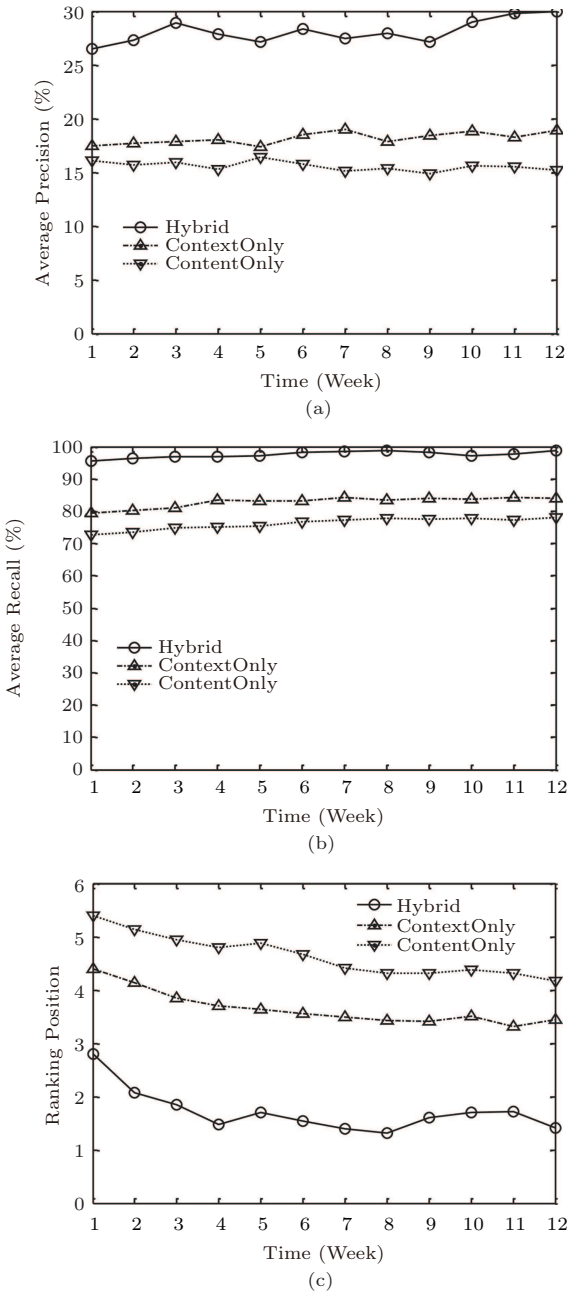


Fig.9. Comparison results with/without context and content cues on a 12-week user study ($k = 10$, $\tau_{\mathcal{E}} = 20\%$, "AdaptDecay").

8 Limitations

For PhotoPrev, content extraction from image is still a challenging problem. To parse an input image, we match the visual objects between the input image and the images in a large database. However, human annotation can be ambiguous. Smaller objects are usually overwhelmed by the labeling of larger objects, which affect the quality of extracted content cues. If images in the database are annotated with object category or the matching is semantically meaningful, we can easily transfer the labels. Otherwise, there are some failure cases including the misclassification of mountain into field, window into wall, and so on.

9 Conclusions

In this work, we proposed a method to automatically construct an adaptive and evolutive context and content memory based on users' personal photo collections, supporting users' photo revisitation by keyword-based queries on smart phones. The proposed method is evaluated by an experiment on a large synthetic dataset and a 12-week user study. Our experimental results show that it can adapt to the user's revisit habit and offer a simple yet effective solution using human memory cues. As future work, we would like to deal with context and content ambiguity considering confusion and error during memory construction.

References

[1] Naaman M, Song Y J, Paepcke A *et al.* Automatic organization for digital photographs with geographic coordinates. In *Proc. the 4th ACM/IEEE Joint Conference on Digital Libraries*, June 2004, pp.53–62.

[2] Naaman M, Harada S, Wang Y *et al.* Context data in geo-referenced digital photo collections. In *Proc. the 12th ACM International Conference on Multimedia*, Oct. 2004, pp.196–203.

[3] Cao L, Luo J, Kautz H *et al.* Annotating collections of photos using hierarchical event and scene models. In *Proc. the 21st IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

[4] Joshi D, Luo J. Inferring generic activities and events from image content and bags of geo-tags. In *Proc. the 7th International Conference on Content-Based Image and Video Retrieval*, July 2008, pp.37–46.

[5] Viana W, Filho J B, Gensel J *et al.* PhotoMap — Automatic spatiotemporal annotation for mobile photos. In *Proc. the 7th Int. Symp. Web and Wireless Geographical Information Systems*, Nov. 2007, pp.187-201.

[6] Viana W, Hammiche S, Villanova-Oliver M *et al.* Photo context as a bag of words. In *Proc. the 10th IEEE International Symposium on Multimedia*, Dec. 2008, pp.310-315.

[7] Crandall D, Felzenszwalb P, Huttenlocher D. Spatial priors for part-based recognition using statistical models. In *Proc. the 18th IEEE Conference on Computer Vision and Pattern Recognition*, June 2005, pp.10-17.

[8] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In *Proc. the 18th IEEE Conference on Computer Vision and Pattern Recognition*, June 2005, pp.886-893.

[9] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. In *Proc. the 21st IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

[10] Felzenszwalb P F, Huttenlocher D P. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005, 61(1): 55-79.

[11] Hu J, Pei J, Tang J. How can I index my thousands of photos effectively and automatically? An unsupervised feature selection approach. In *Proc. the 14th SIAM International Conference on Data Mining*, Apr. 2014, pp.136-144.

[12] Zhou W, Li H, Lu Y *et al*. Encoding spatial context for large-scale partial-duplicate web image retrieval. *Journal of Computer Science and Technology*, 2014, 29(5): 837-848.

[13] Shotton J, Winn J, Rother C *et al*. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 2009, 81(1): 2-23.

[14] Hu S, Chen T, Xu K *et al*. Internet visual media processing: A survey with graphics and vision applications. *The Visual Computer*, 2013, 29(5): 393-405.

[15] Frome A, Singer Y, Malik J. Image retrieval and classification using local distance functions. In *Proc. Neural Information Processing Systems*, Dec. 2006, pp.417-424.

[16] Russell B C, Torralba A, Liu C *et al*. Object recognition by scene alignment. In *Proc. Neural Information Processing Systems*, Dec. 2007, pp.1241-1248.

[17] Russell B C, Torralba A, Murphy K P *et al*. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 2008, 77(1/2/3): 157-173.

[18] Liu C, Yuen J, Torralba A. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(12): 2368-2382.

[19] Liu C, Yuen J, Torralba A. Sift flow: Dense correspondence across different scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(5): 978-994.

[20] Cao W, Liu N, Kong Q *et al*. Content-based image retrieval using high-dimensional information geometry. *SCIENCE CHINA Information Sciences*, 2014, 57(7): 1-11.

[21] Gllavata J, Ewerth R, Freisleben B. Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. In *Proc. the 17th International Conference on Pattern Recognition*, Aug. 2004, pp.425-428.

[22] Chen X, Yuille A L. Detecting and reading text in natural scenes. In *Proc. the 17th IEEE Conference on Computer Vision and Pattern Recognition*, June 2004, pp.366-373.

[23] Ye Q, Huang Q, Gao W *et al*. Fast and robust text detection in images and video frames. *Image and Vision Computing*, 2005, 23(6): 565-576.

[24] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform. In *Proc. the 23rd IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp.2963-2970.

[25] Lee J, Lee P, Lee S *et al*. AdaBoost for text detection in natural scene. In *Proc. the 12th International Conference on Document Analysis and Recognition*, Sept. 2011, pp.429-434.

[26] Matas J, Chum O, Urban M *et al*. Robust wide baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 2004, 22(10): 761-767.

[27] Neumann L, Matas J. Real-time scene text localization and recognition. In *Proc. the 25th IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp.3538-3545.

[28] Zhang X, Lin Z, Sun F *et al*. Transform invariant text extraction. *The Visual Computer*, 2013, 30(4): 401-415.

[29] Chen T, Chen M, Tan P *et al*. Sketch2Photo: Internet image montage. *ACM Transactions on Graphics*, 2009, 28(5): Article No. 124.

[30] Lee Y, Zitnick C L, Cohen M F. ShadowDraw: Real-time user guidance for freehand drawing. *ACM Transactions on Graphics*, 2011, 30(4): Article No. 27.

[31] Ellis H C. Fundamentals of Human Memory and Cognition (3rd edition). William C. Brown Press, 1983.

[32] Rubin D C, Wenzel A E. One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 1996, 103(4): 734-760.

[33] Tulving E. What is episodic memory? *Current Directions in Psychological Science*, 1993, 2(3): 67-70.

[34] Wiggs C L, Weisberg J, Martin A. Neural correlates of semantic and episodic memory retrieval. *Neuropsychologia*, 1999, 37(1): 103-118.

[35] Ding Y, Li X. Time weight collaborative filtering. In *Proc. the 14th ACM International Conference on Information and Knowledge Management*, Oct. 2005, pp.485-492.

[36] Fagin R, Lotem A, Naor M. Optimal aggregation algorithms for middleware. In *Proc. the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, May 2001, pp.102-113.

[37] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. the 18th International Conference on Machine Learning*, June 28–July 1, 2001, pp.282-289.

Li Jin received his Bachelor's degree in computer science and technology from Xidian University, Xi'an, in 2012. He is currently a Ph.D. candidate in the Department of Computer Science and Technology, Tsinghua University, Beijing. His research interests include context-aware data management and context-based information refinding.

Gang-Li Liu received his Bachelor's and Master's degrees in communication engineering from the PLA University of Science and Technology, Nanjing, in 2003 and 2006, respectively. He is currently a Ph.D. candidate in the Department of Computer Science and Technology, Tsinghua University, Beijing. His research interests include context-aware data management and context-based information refinding.

Liang Zhao received her Bachelor's degree in computer science and technology from Xidian University, Xi'an, in 2011. She is currently a Ph.D. candidate in the Department of Computer Science and Technology, Tsinghua University, Beijing. Her research interests include context-based information refinding and user revisit interest prediction.

Ling Feng is a professor of computer science and technology at Tsinghua University, Beijing. Her research interests include context-aware data management toward ambient intelligence, knowledge-based information systems, data mining and warehousing, and distributed object-oriented database management systems. She has published more than 150 scientific articles in high-quality international conferences or journals, and received the 2004 Innovational VIDI Award by the Netherlands Organization for Scientific Research, the 2006 Chinese ChangJiang Professorship Award by the Ministry of Education, and the 2006 Tsinghua Hundred-Talents Award.