Hindawi Publishing Corporation
Advances in Multimedia
Volume 2008, Article ID 592690, 16 pages
doi:10.1155/2008/592690

Research Article

On the Design and Exploitation of User's Personal and Public Information for Semantic Personal Digital Photograph Annotation

Supheakmungkol Sarin,[1] Toshinori Nagahashi,[2] Tadashi Miyosawa,[2] and Wataru Kameyama[1]

[1] *Graduate School of Global Information and Telecommunication Studies, Waseda University, 1011 Okuboyama, Nishi-Tomida, Honjo-shi, Saitama-ken 367-0035, Japan*
[2] *Research and Development Division, Seiko Epson Corporation, 80 Harashinden, Hirooka, Shiojiri-shi, Nagano-ken 399-0785, Japan*

Correspondence should be addressed to Supheakmungkol Sarin, mungkol@fuji.waseda.jp

Received 8 September 2007; Revised 18 February 2008; Accepted 10 March 2008

Recommended by Q. Sun

Automating the process of semantic annotation of digital personal photographs is a crucial step towards efficient and effective management of this increasingly high volume of content. However, this is still a highly challenging task for the research community. This paper proposes a novel solution. Our solution integrates all contextual information available *to* and *from* the users, such as their daily emails, schedules, chat archives, web browsing histories, documents, online news, Wikipedia data, and so forth. We then analyze this information and extract important semantic terms, using them as semantic keyword suggestions for their photos. Those keywords are in the form of named entities, such as names of people, organizations, locations, and date/time as well as high frequency terms. Experiments conducted with 10 subjects and a total of 313 photos proved that our proposed approach can significantly help users with the annotation process. We achieved a 33% gain in annotation time as compared to manual annotation. We also obtained very positive results in the accuracy rate of our suggested keywords.

1. INTRODUCTION

The low cost of today's digital cameras and digital storage devices, combined with the rapid adoption of broadband Internet connectivity, has given consumers greater control over their digital photo collections than ever before. Consumers can now capture and store thousands of their digital photographs on their personal computers. They can also speedily share them with their friends over the Internet. However, with the rapid growth of personal digital photos, the complexity and difficulty in archiving, searching, and browsing photographs have also proportionately increased. In most case, users can only organize their digital photographs with the tree structure of the file system they are using, which is quite limited and unnatural. Hence, users cannot fully enjoy their photos because the value of the photos depends largely on how they can effectively and efficiently access, manage, and share them. Until today, there were no complete real-world solutions to this growing problem.

These above-mentioned problems are due to the lack of rich metadata associated with photos. Annotation is one of the key solutions to enable better access to digital photographs. In other words, users need to provide contextual metadata (meaningful descriptions) to each of their photograph files. This would allow them to find their photos by searching using more abstract information instead of the file or directory names. However, this annotation process is tedious and time consuming for users. Factor in the need to annotate hundreds or thousands of photos, and the task quickly becomes unrealistic for the average user to conduct or keep up with. Research shows that although people would like their photo albums to be organized, many do not label more than only a few, or they do not invest the effort to label their photos at all [1]. Therefore, most photos are poorly annotated or just retain the numerical file names that the camera defaults to.

Various research efforts on how to annotate images have been going on actively in the last decade. On one hand, there are techniques to extract relevant metadata directly from

image content which include color/texture extraction, object identification, face detection/recognition, content-based categorization, and so forth. In 2000, Smeulders et al. published a comprehensive survey of these techniques [2]. However, these content-based technologies hold limited value as they are often inaccurate and too vague to accurately represent the interpretation of each individual. Other methods involve designing a better graphical annotation interface in order to allow users to easily input contextual metadata manually. In addition to this, there are approaches that depend on users' collaboration. One of them is an ESP game-like approach that is gaining popularity by using the power of anonymous volunteers to help manually label the photos over the web [3]. This concept is also adopted by Google Image Labeler [4]. However, these collaborative approaches require consistent participation from users, consuming both their time and energy. Unfortunately, this approach will never work for annotating personal photos, which often require private knowledge and contextual information of the owner's ambient environment and application of his or her personal interpretation of the environment and moment. Other methods try to use both content and context information such as that of Tuffield et al. [5]. However, the work is still very primitive and the authors only limit to a few kinds of contextual information. Datta et al., recently, produced a detailed survey paper of the progress report in the field from the year 2000 [6]. We will also elaborate more about the closely related techniques to ours in the related work section.

In our study, we look at the problem by asking the following question: *how can we generate semantic metadata for photos without requiring the owner to manually input the data?*

We answer this question by proposing *to use the maximum amount of contextual information about the photos that are available from and to the users*. Information from the photo owners, such as their emails, schedules, web browsing histories, and files and information available to the owners, such as news and encyclopedia are the focus of this study. We introduce a practical implementation paradigm to leverage the above-mentioned information which serves as personalized and contextual metadata to suggest back as the semantic metadata for the photos. We do this by assuming that the exact location information is available for every captured photo based on the current trend in geophotography. We use this location data in addition to timestamp data of the captured photo as "information filters" for relevant contextual information of that photo. By applying information extraction and retrieval techniques to the filtered contextual information, our system can suggest accurate semantic keywords to each photograph. Moreover, we propose to use named entities, such as the names of people and organizations, to represent the exact semantic meaning of the photos in addition to the high frequency terms.

We have designed and implemented a prototype of our proposed system. We have also performed the experiments to verify the effectiveness and accuracy of the system. Results show that users are able to annotate their photos significantly faster using our proposed system. We have also obtained an encouraging rate of accuracy.

2. PROPOSED APPROACH: LEVERAGING CONTEXT TO BRIDGE SEMANTIC GAP

2.1. Nature of personal digital photographs

An image or photograph can mean different things to different people. An image itself has no intrinsic meaning. Instead, meaning is bestowed upon the image by the viewer. Personal digital photographs have very different characteristics when compared with other types of images, such as those found in museums or web image collections. Usually a user's personal digital photos reflect their daily activities. The information from one's daily life is the ideal resource to be used to extract the semantic information needed to describe photos taken on a particular day or within a short interval.

2.2. Gathering contextual information

Many of us use computers both at home and at work. We use them to prepare or consult our schedules, read or write emails, surf the Internet, and get or share information with family, friends, and colleagues via various Internet services such as chats, forums, and blogs.

In a typical scenario, suppose that we are going for a trip, we might have planned this ahead in our schedules. Before leaving, we book a hotel room online, find the nearest public transportation, and look for general information about the place we are to visit, such as weather, culture, main attractions, and related news. We might use encyclopedia and tourism websites, online news, and other sources. We might also email or chat with our friends and family about our upcoming trip. On the spot, we take lots of photos while we enjoy the trip. Upon returning, we share the photos as well as our impressions about the places with our friends and family via the Internet services mentioned earlier. This is often very useful information, as it comes from a user's direct personal interpretation of the photos (via their schedules, emails, chats, etc.) as well as from the other information they are processing from their environment (such as Wikipedia, tourism websites and online news websites, to name a few). These sources of information are important because what occurs in the ambient environment will add both direct and indirect effect to a user's episodic memory. When looking for photos later, users are very likely to use the same keywords that they use in personal documents and in describing experiences in their ambient environment. We categorize these sources of information into two types.

(1) *Personal information* refers to available contextual information from users such as schedules, notes, emails, chats, web browsing histories, and all other documents residing in their computer or computers. These types of information link to users directly and personally.

(2) *Public information* refers to contextual information that users consume freely or very cheaply such as local news, world news, encyclopedia information,

tourism information, and other information from public repositories that are available online. These types of information link to users directly or indirectly.

2.3. Using time and location as photograph filters

As mentioned earlier, the *personal* and *public information* is readily or cheaply available, which provides for some huge advantages. However, a method is needed that allows us to distinguish which subset of the acquired information best represents the context of a captured photo. To do this, we consider the time and location of each photo as the key filters, because this information serves as the basic contextual metadata of the photo.

All digital cameras now provide time information. A timestamp indicating exactly when the photo was captured is embedded in each photo file itself. In addition, most camera phones can infer a rough location from GPS or cell ID information. It is likely that all new cameras will eventually be equipped with location capturing systems. Additionally, most digital photographs support location data in addition to time information. This data can be stored in the form of a coordinate set (longitude and latitude) in the EXIF header [7] of every photograph. There are documented trends as far as providing free location information database to the general public. For instance, geonames [8] provides free geodata such as geographical names and postal codes to the public, and its database contains over 8 million entries of geographical names within 2.2 million are cities and villages. Geonames's website boasts many features, including conversion from GPS coordinate set to nearby location. Consequently, there is no problem as far as translating a GPS coordinate set into an exact location name. As a result of services such as these, we will be able to obtain two key filters, namely, timestamp and location, without much effort in the near future.

Based on the above facts and hypothesis, knowing the exact time and location where a photo was taken can be used to extract the subset of personal and public information from a user's pre-scene (before going) and post-scene (after going) that strongly relates to a photo or group of photographs. By applying some natural language processing techniques to this obtained information, we will be able to extract important representative keywords and suggest them to users for their validation.

2.4. Extracted keywords

We identify two classes of keywords to be extracted.

(1) *Named entity keywords* refer to strong and exact proper noun identifications found in the relevant files. To generate this type of keywords, we employ computational linguistic techniques to intelligently parse documents and discover named entity (NE) information. In our case, we would like to get the important episodic memory information such as *dates, names of people, location names, and organization names* .

(2) *Statistical keywords* refer to terms that appear very frequently in the relevant files and that can be used to represent these files.

Figure 1 illustrates our concept.

3. SYSTEM DESIGN AND IMPLEMENTATION

We have designed and implemented a prototype of our system. The overall architecture of our system is depicted in Figure 2 The following is the step-by-step explanation of the annotation process with our semiautomatic annotation system.

(1) Users begin by choosing the photo that they would like to annotate. It is assumed that these photographs are embedded with date/time and location information. In our case, the file name of each photo contains location name.

(2) The extracted date/time and location are used as key filters to search for related sources from user's computers including their personal and public information. Google Desktop Search (GDS) returns to us the relevant files from its index.

(3) Relevant files to the photo with respect to time and location are sent to the named entity extraction module. In return, NEs from the relevant files with respect to their categories, namely, date, location, people's name, organization will be output. In addition, those output NEs are ranked by their frequencies of occurrence.

(4) In the same manner as the previous step, all the relevant files related to the photo are sent to statistical keyword extraction module. This module processes the term ranking and outputs the top keywords ranked by their frequency of occurrence in the document sources.

(5) In this step, metadata (NEs and statistical keywords) found in steps (3) and (4) are presented to the users. Top suggested keywords of each category are shown in their respective fields of the interface. Users may consult more keywords by clicking on the *magnifying icon* of each field. Finally, users validate the metadata candidates (they can always edit or augment the metadata if necessary).

(6) All the metadata validated by users are converted to MPEG 7 MDS format and are sent to our exist XML database.

All detailed processes are described as follows.

3.1. Data acquisition

Personal information of a user resides in their computers. Currently, there is a tremendous interest in desktop search. Desktop search engine software can index and search files on a single computer or across multiple networked computers. The world's top software companies such as Google, Yahoo!,

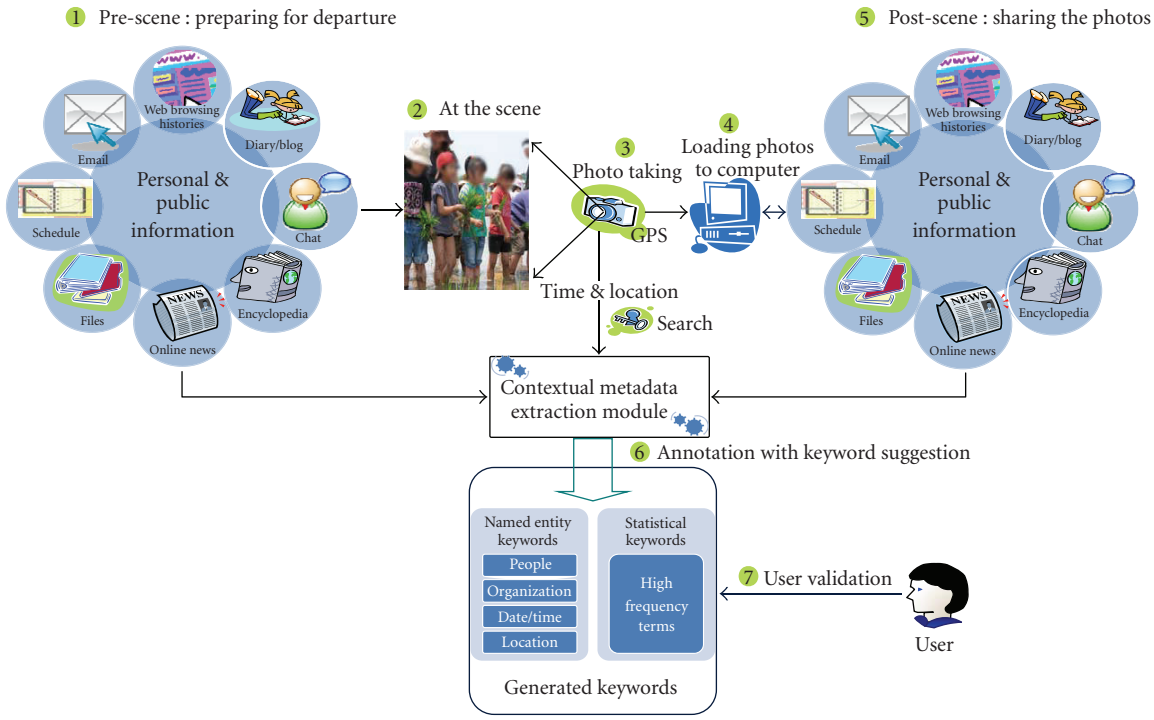


Figure 1: Overall view of the concept.

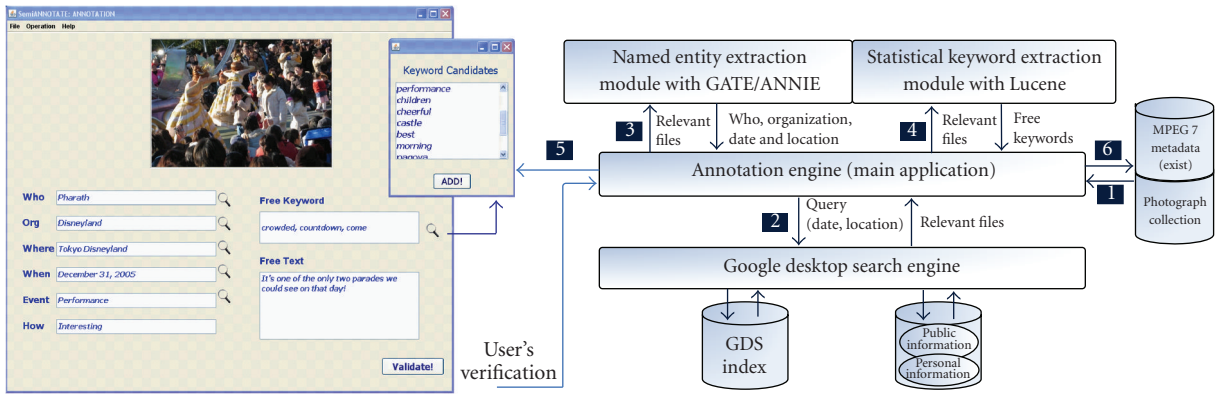


Figure 2: System architecture of the implemented prototype.

and Microsoft offer their proprietary versions of the desktop search application. Lu et al. have a comprehensive analysis about the various kinds of desktop search software currently available and their performance metrics [9].

Google Desktop Search (GDS) [10] is among the most popular desktop search applications. GDS manages and indexes files found on personal computers. These files include email, schedule, web browsing history from Internet Explorer and Mozilla Firefox, office documents in the open document and Microsoft Office formats, memo, PDF, instant messenger transcripts from AOL, Google, MSN, Skype, and several multimedia file types. GDS includes plugins for different file formats that allow one to index and search through the contents of those local files. Google Desktop's email indexing feature is also integrated with Google's web-based email service called Gmail. GDS performs all tracking, cataloging, and indexing entirely independently of the Windows caching of Internet pages. Therefore, should a user delete his temporary Internet files, cache, and cookies, a record of the data is maintained by the GDS program. This means GDS caches all HTML Internet pages visited. Additionally, should a single web page have been visited repeatedly, the Google Desktop Search will store cached copies of all of these pages, giving exact information on what was presented to the browser on each visit. In addition, GDS is designed to index and retrieve user-created data only. Consequently, it does not index system-related files such as Microsoft Windows system files. Files stored within the default Windows directory, within the Recycle bin, or those that are invisible are not be indexed. They are excluded

from indexing, increasing the efficiency of the program [11]. Another feature of GDS is called Search Across Computers. This feature enables us to search our files and viewed Web pages across all of our computers. For example, one can find files that he or she edited on the desktop from his/her laptop. To activate this feature, a Google account is needed and the GDS program must be installed on each computer [12].

With these above-mentioned qualifications, we decide to choose GDS as our data acquisition tool. This enables us to access all of the personal information residing on the user's computer. In our case, to make it simple, we also make public information available to GDS so that it can index this together with personal data. To do so, we download news and encyclopedia data from the Internet, and maintain them in the local directories on the user's personal computer. We consider the following online public repositories as the *public information* to be integrated:

(1) news: MDN Mainichi Daily News [13], The Asahi Shimbun [14] (in English and in duration of two-year time),

(2) encyclopedia: English Wikipedia [15].

The news pages are downloaded via a tool called HTTrack [16]. The tool is configured to download only printer-friendly version of its HTML pages to minimize the tasks needed to clean up the unnecessary information in the page such as advertisements, pictures, flash media. GDS is integrated into our system via its Java API, which is available from the SourceForge website [17]. Figure 3 summaries the process.

3.2. Relevant files generation

Google Desktop Search also serves as our search tool for relevant indexed sources to date and location. This allows us to leverage Google's search technology. GDS is designed to narrow search space to areas that are more likely to contain documents stored by the user rather than files used to operate and maintain the computer. We define three patterns of queries to GDS to enable both exact and loose query in case the number of exact relevant sources is limited. We limit the maximum size of the result set to 100 in order to assure the quality of our metadata and the efficiency of the approach by maintaining both relevancy and computing performance. Figure 4 shows the process in generating relevant files. Algorithm 1 is used to retrieve relevant contextual information for the photos from public and personal information resources.

3.3. Keywords generation

3.3.1. Named entity keywords

To get this type of keywords from relevant sources, information extraction techniques are needed. For this purpose, we integrate the General Architecture for Text Engineering (GATE) [18], a mature open source text engineering platform, into our system. GATE comes with A Nearly New Information Extraction (ANNIE) engine, a robust information extraction engine based on finite state algorithms. ANNIE depends on a number of language processing tools to do named entity extraction range from unicode tokenizer, sentence splitter, part-of-speech tagger, gazetteers, semantic tagger to name matcher and pronominal coreferencer. We introduce some linguistic resources specific to our situation such as company names, city names, people's names. We also developed an NE sorting and ranking module associated with the GATE/ANNIE module. Top 20 NE keywords are generated for each category of keywords. Figure 5 depicts the process of named entity keywords extraction. We describe each element as follows.

(i) The tokenizer splits the text into very simple tokens such as numbers, punctuation, and words of different types.

(ii) The gazetteer lists used are plain text files, with one entry per line. Each list represents a set of names, such as names of cities, organizations, days of the week.

(iii) The sentence splitter is a cascade of finite-state transducers which segments the text into sentences. This module is required for the tagger. The splitter uses a gazetteer list of abbreviations to help distinguish sentence-marking full stops from other kinds.

(iv) ANNIE's semantic tagger is based on the JAPE language. It contains rules which act on annotations assigned in earlier phases, in order to produce outputs of annotated entities.

(v) The name matcher module adds identity relations between named entities found by the semantic tagger, in order to perform coreference. It does not find new named entities as such, but it may assign a type to an unclassified proper name, using the type of a matching name.

(vi) The pronominal coreference module performs anaphora resolution using the JAPE grammar formalism.

(vii) Named Entity Sorter ranks and sorts the found NEs according to their frequencies of appearance and their category.

3.3.2. Statistical keywords

Google Desktop Search is a closed technology of Google. We cannot fully configure and program it to analyze its index. Therefore, we also need a tool to index those related documents in order to perform other kinds of keyword extractions. Lucene [19] is a good tool to use to accomplish this. Lucene is an open source information retrieval library. At the core of Lucene's logical architecture is the idea of a document containing fields of text. This flexibility allows Lucene's API to be independent of file formats. Texts from PDFs, HTML, Microsoft Word documents, and many others can all be indexed as long as their textual information can be extracted. In our case, we index all the relevant files in the different formats by the Lucene module that we developed

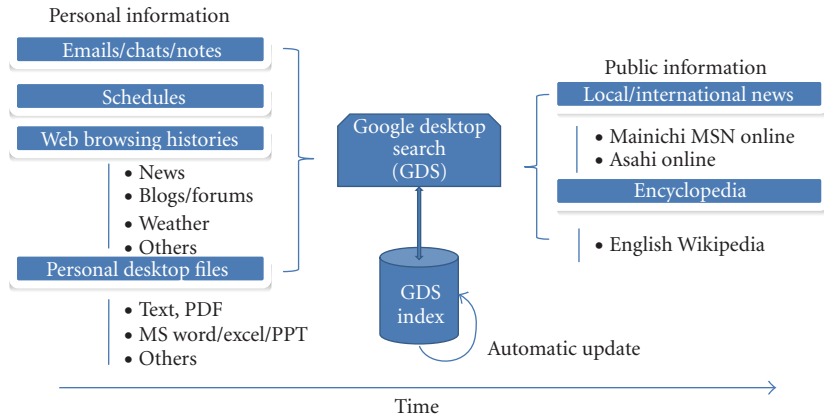


Figure 3: Data acquisition of personal and public information with Google Desktop Search.

Require: *gds_index*, *date*, *location*
Ensure: *relevantFiles* = *generateRelevantFiles*(*gds_index*, *date*, *location*)
1: *resultSet*1 ⇐ *gds_index.query*(*date.getMonthYearDay*(), *location*)
2: *resultFiles* ⇐ *resultSet* 1
3: **if** *relevantFiles.getSize*() < 100 **then**
4: *resultSet*2 ⇐ *gds_index.query*(*date.getMonthYear*(), *location*)
5: *relevantFiles* ⇐ *relevantFiles.add*(*resultSet*2)
6: **if** *relevantFiles.getSize*() < 100 **then**
7: *resultSet*3 ⇐ *gds_index.query*(*date.getYear*(), *location*)
8: *relevantFiles* ⇐ *relevantFiles.add*(*resultSet*3)
9: **end if**
10: **end if**

Algorithm 1: Generate relevant files.

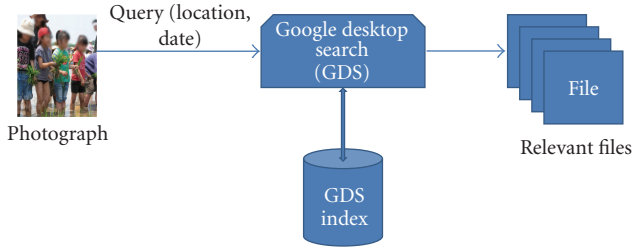


Figure 4: Process of generating relevant files to the photo with location and time as event filter.

using the Lucene's Java API. With this index, we calculate the statistics of each term to find the most frequent terms in the document collection that can be used as representative terms. Top 30 keywords are then generated for each photo. The following shows how we calculate the frequency of each term.

Let

(i) TF (i, j) be the number of occurrences of term $t(i)$ in document $d(j)$,

(ii) DL (j) be document length or the total of term occurrences in document $d(j)$,

(iii) n be the number of relevant sources.

A simple count is too crude because a term that occurs the same number of times in a short document is likely to be more valuable than in a long one. Therefore, we employ a simple adjustment based on the length of document. Hence, the term frequency is computed as the follows:

$$\mathrm{TF}_n(i) = \sum_{j=1}^{n} \frac{\mathrm{TF}(i,j)}{\mathrm{DL}(j)}. \quad (1)$$

Figure 6 illustrates the process in statistical keyword extraction.

3.4. Annotation GUI and metadata coverage

In our annotation GUI, we have correspondent text field for each of the categories of keywords. Below is the description of each one of them.

(i) *Who* refers to people's name.

(ii) *Org.* refers to organization name.

(iii) *Where* refers to location name.

(iv) *When* refers to date/time.

(v) *Free Keywords* refers to statistical keywords.

Among generated NEs and statistical keywords, by default, the first top NE is inserted in the *Who* and *Org.* fields while

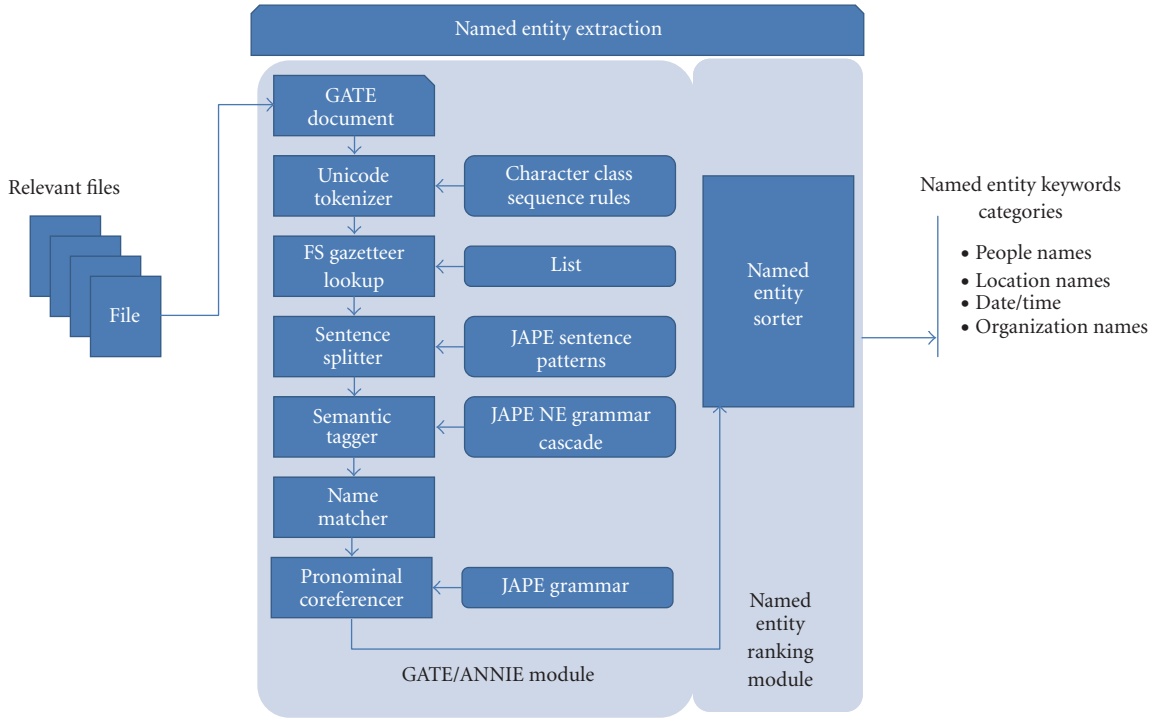


Figure 5: Named entity keyword extraction process.

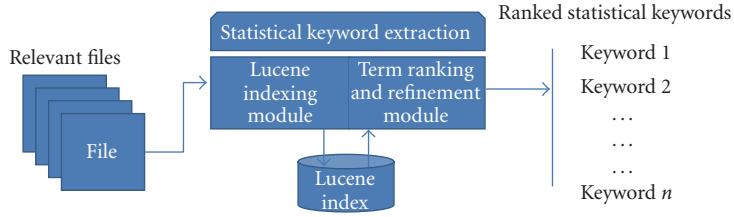


Figure 6: Statistical keyword extraction process.

3 statistical keywords are inserted in the *Free Keywords* field of the annotation interface. *When* and *Where* fields are also filled, respectively, with time and location of the photo. Users can always edit those default keywords if necessary. In addition to these automatically generated keywords, we also have other categories of keywords in our interface. They include the following.

(i) *Event* refers to reasons about the photos. We prepare some preset values for it with a list of events such as birthday, wedding, meeting, graduation, festival, New Year that users can select from or add their own keywords.

(ii) *How* refers to actions or emotions about the photos.

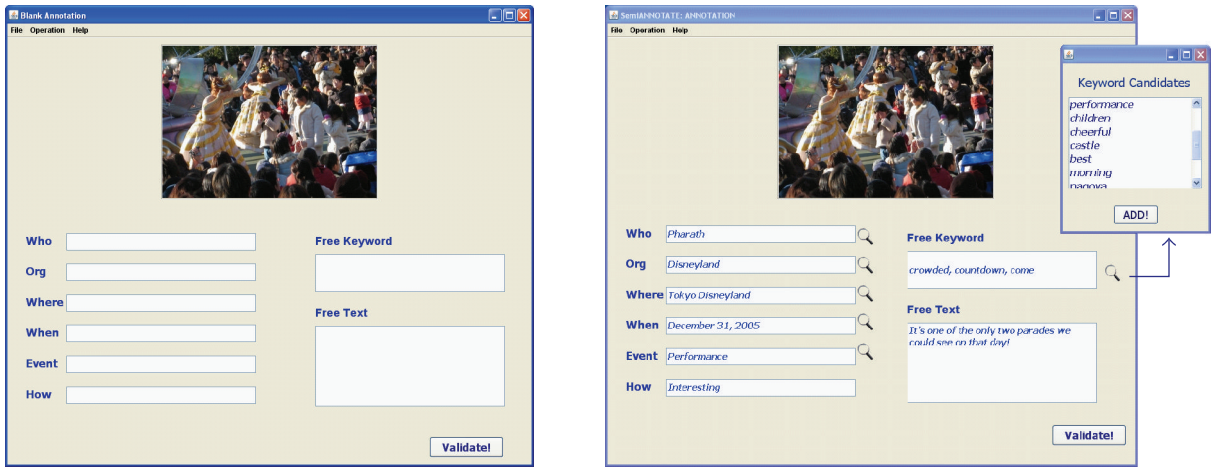
(iii) *Free Text* refers to free text description about the photos.

We introduced these additional categories to improve the semantic integrity of our metadata for the retrieval task. Even though *Event* and *How* are not suggested by the current system, we believe that these keywords can be covered by the statistical keywords that we generate. Therefore, we can cover all of the related questions about photos including the *W5H1* (Who, What, Where, When, Why, and How) questions (*What* could also be found in the statistical keywords). Please refer to Figure 7(b) for our annotation interface.

3.5. Metadata format and storage database

Contrary to Dublin core [20] which aims at simplicity, MPEG-7 [21] provides ways to give rich description for audio-visual media. Since our work focuses on semantic metadata about the photo, MPEG-7 element set is the best choice. In our case, we extended the *Structured Annotation* Basic Tool of MPEG-7 multimedia description schemes (MDS) [22] to adapt and include all the categories of metadata extracted.

Since our MPEG-7 metadata is XML based, we also choose an XML native database to store the photo metadata in order to enhance the retrieval capabilities (search and browse). We choose eXist database for this purpose. eXist is an Open Source native XML database featuring efficient,



(a) (b)

Figure 7: (a) Blank annotation interface, (b) annotating interface with keywords suggestion feature.

index-based XQuery processing, automatic indexing, extensions for full-text search, XUpdate support and tight integration with existing XML development tools [23].

4. EMPIRICAL EVALUATIONS

4.1. Validation goals

We investigate the performance of our system on two grounds:

(1) The time difference between manual annotation and annotation by our proposed system using the built-in keyword suggestion features.

(2) The accuracy of our proposed named entity keywords and statistical keywords by calculating their acceptable hit rates.

4.2. Participants and data sets

4.2.1. Subjects

We were able to recruit ten subjects for the experiments of our system. All subjects were computer science students at the Graduate School of Global Information and Telecommunication Studies of Waseda University. They are all familiar with computers; they use and work with computers in their daily lives. Three of the subjects were women and seven were men.

4.2.2. Personal photographs

Each subject was asked to provide more than 30 personal photographs which had been taken over a period of six months. Photos are taken from events such as sightseeing, friend-gatherings, dinner parties, picnics, and so forth. Each subject provided photos for an average of 5 events. Each event had about 5 photos. We gathered 313 photographs in all.

Subjects were asked to install Google Desktop Search (GDS) and activate it each time they used their computers. Though GDS has its own cache index file system as described in Section 3.1, the subjects were requested not to delete any of the files on their computers. This was required so that we can generate links to original files during the relevant files generation process. Subjects were also required to install our prototype system on their computers.

As mentioned in Section 3.1, we manually downloaded the news from online repositories and Wikipedia. We then bundled this data into one single folder named *public_information* and asked the subjects to save it on their computers. Google Desktop Search was then configured to include this folder into its index.

4.3. Experiment process

First, in order to enable location information for each photo, we asked the subjects to label their own photos with the exact location name as the file name of the photo. To do this, we provide a drag-and-drop interface where subjects can easily input the location name on their photo(s).

The experiment is three-part process. The first two parts are for time evaluation and the third one is to measure the accuracy. First, subjects are expected to annotate their own photos manually. Second, subjects were asked to annotate their photos using our proposed prototype system with keyword suggestion features. Between the two parts of the experiment, we leave a gap of 2 to 3 days so that subjects have time to forget their previously input keywords. This is done to avoid the influence of a subject's memory about the keywords of the photos that they have input into the system during the first step. Users were asked to input at least one keyword to the *Who* and *Org.* fields. They have to input at least three keywords in the *Free Keyword* field. Lastly, subjects were requested to judge the accuracy of the automatically generated keywords for each photos that we saved into files before we performed the second step.

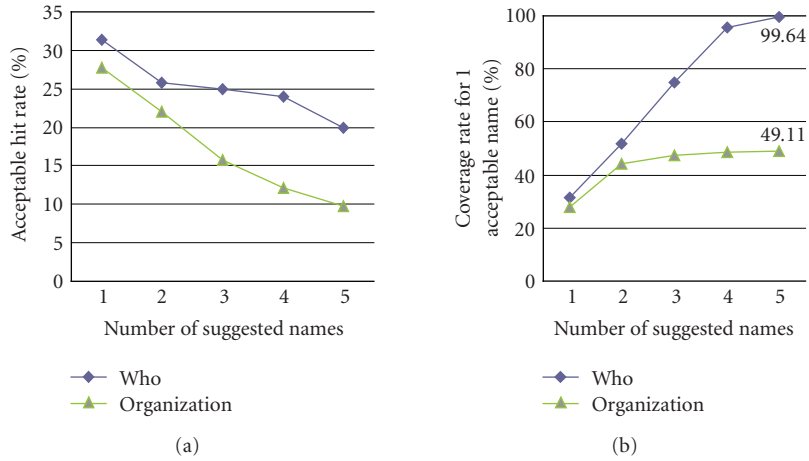


Figure 8: (a) Acceptable Hit Rate of Who (people's name) and Org. (organization names) keywords, (b) Coverage Rate for 1 acceptable keyword of Who (people's name) and Org. (organization names).

Please also note that we performed the experiment on the subjects' own computers, using their own contributed photos. Therefore, the timing varies depending on the configuration of their PCs. More details about the three parts of the experiment follow.

4.3.1. Manual annotation

Users begin the experiment by manually annotating their photos with a blank interface. A blank interface is similar to the interface of our proposed system. It has all the fields for every category of keywords. However, the only difference is that there is no suggestion feature on this interface. Each text field represents a category of keywords accordingly. Thus subjects have to manually input the annotation keywords to each text field. Annotation time is recorded for each photo. Figure 7(a) shows our blank annotation interface.

4.3.2. Annotation with keyword suggestion features

In this step, subjects annotate their photos with the help of our system. Top keywords of each field are shown in the respective text field. Subjects can consult other less ranked keywords by clicking on the magnifying icon and selecting from a drop-down list of suggested terms. At any time, subjects can modify the suggested keywords or add their own keywords if they find it necessary. Figure 7(b) shows our annotation interface with the keyword suggestion features.

It is noted that in case no relevant file is found, the top NE keywords and statistical keywords found in the total index will be suggested. In the same way as in the previous task, we record the annotation time of each photo. It is also noted that at the beginning of this step, for each photo, we automatically generate the following: 30 free keywords, 5 person names, and 5 organization names. We then save these to a file for the last step of the experiment (keyword judging).

4.3.3. Keywords judging

In this step, we asked the subjects to work on the automatic keyword candidates of each field that we have generated. Subjects had to identify all the *acceptable keywords* of each field manually. Acceptable keywords refer to all the keywords that relate to the photo and are appropriate as keywords to describe or recall the photo.

4.4. Results and discussion

4.4.1. Experimental results and analysis

Accuracy

We evaluate the accuracy and the coverage of suggested keywords by using the following formulas:

$$\text{Acceptable Hit Rate}(p,k) = \frac{\sum_{j=1}^{p}\sum_{i=1}^{k} \mathrm{H}_j(i)}{p \times k},$$
$$\text{Coverage Rate}(p,k,n) = \frac{\sum_{j=1}^{p}\sum_{i=1}^{k} \mathrm{H}_j(i)}{p \times n}, \tag{2}$$

where

(i) p is the total number of photos,

(ii) k is the number of suggested keywords,

(iii) n is the number of acceptable keywords expected,

(iv) $\mathrm{H}_j(i)$ is the hit function of keyword i to photo j:

(a) $\mathrm{H}_j(i) = 0$ if the keyword is not acceptable,

(b) $\mathrm{H}_j(i) = 1$ if the keyword is acceptable.

Figure 8(a) shows that the acceptable hit rates of proposed names of people and organization drop gradually from 31% (Who) and 27% (Org.) to 19% and 9%, respectively, when the number of names is suggested from 1 to 5. The first name suggested of both categories can hold about

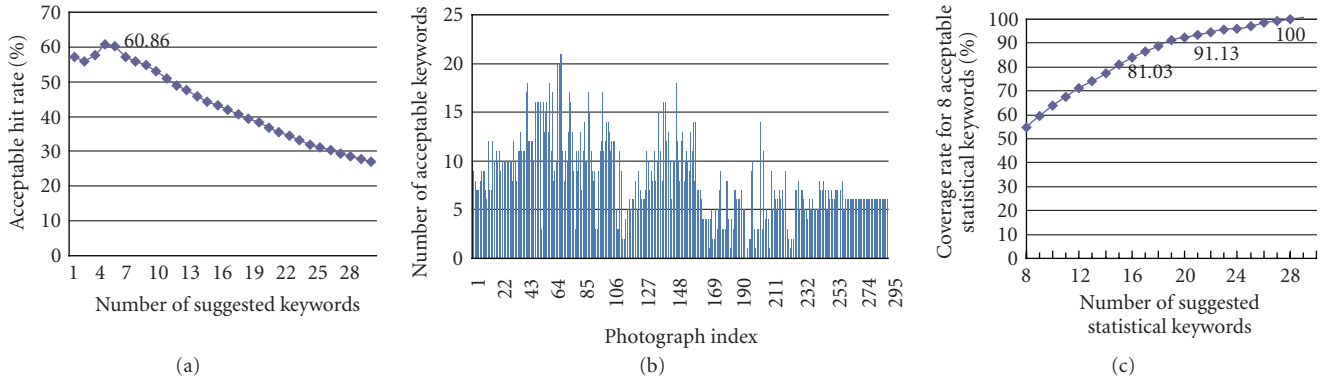


Figure 9: (a) Acceptable Hit Rate of statistical keywords, (b) number of acceptable keywords of each photo, (c) Coverage Rate for at least *8* acceptable keywords of statistical keywords.

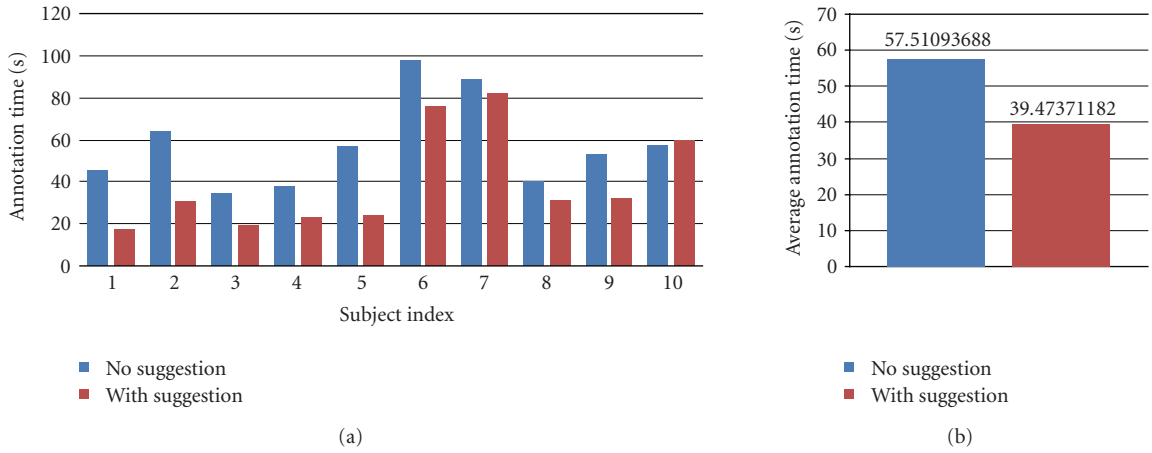


Figure 10: (a) Manual annotation and annotation with keyword suggestion features of each subject, (b) average annotation without and with keyword suggestion features.

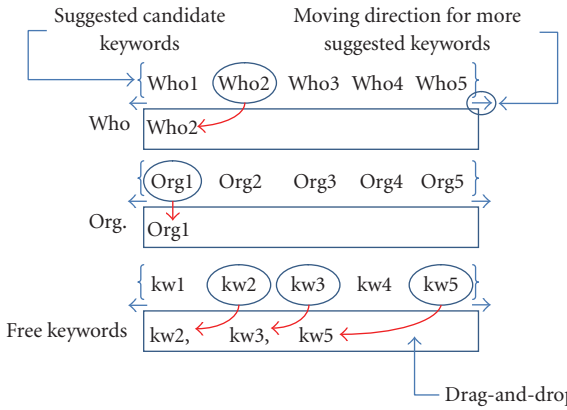


Figure 11: Proposed annotation interface for future implementation.

30% of being acceptable. However, by integrating all the 5 suggested names together, Figure 8(b) suggests that 99% of photos will have at least 1 acceptable person name and about 49% of photos will have at least 1 acceptable name of organization.

Figure 9(a) discusses the accuracy of automatically suggested statistical keywords. We can see that the hit rate reaches its peak level (60%) when we suggest 4 or 5 keywords. This means that we will get 3 acceptable keywords if we suggest 5 keywords to users. This is significant. However, Figure 9(b) shows that, if we automatically suggest 30 keywords, the average number of acceptable keywords of the photos is 8. To further analyze, if we calculate the coverage rate for 8 acceptable keywords to the photos which is the percentage of photos that is correctly suggested by at least 8 acceptable keywords, we come up with the result in Figure 9(c). It shows that to achieve 80%, 90%, or 100% of coverage, we need to supply 15, 20, and 29 keywords, respectively. These results are very encouraging.

Time

We arrive at the following result. Figure 10(a) shows that 9 out of 10 subjects gain benefit from this approach. Figure 10(b) depicts that in average we gain an overall of 33% in annotation time over the traditional manual annotation. This is significant to the users.

Analysis of results

(i) In overall, our approach has allowed us to obtain good accuracy rate and time gain, despite a large diversity of photos and the relative subjectivity of our subjects. However, we should not neglect these influences. For instance, in Figure 10(a), our system cannot overcome the problems of subject number 10. This is due to the fact that the majority of his photos are scenery from trips to different places and include no individual or organization names. In this case, the subject had to take time to edit the incorrectly suggested NE keywords or blank fields (when there are no keywords found by the automated system). In addition, he had to think of new keywords to attribute to his photos manually. We have noted that the type, size, and numbers of files generated by a user most often link to that user's habits. This ultimately influenced the results of this study. We found that the average size of contributed data is always less than 100 KB. Relevant files bigger than this size generally produce more noise. Furthermore, it is the personal information that contributes most to the acceptable keywords. Public information contributes only in the case that the event is a breaking news event or happened in a popular place or time (such as New Year's, Christmas, at the Tokyo Dome). Events such as a simple dinner gathering do not create the same impact. Therefore, we will establish a threshold in order to adjust to these variants.

(ii) Obviously, there is also a strong correlation between the accuracy rate and the annotation time. However, we recognize that designing a better interface can save more time. In our case, subjects have to first click on the magnifying icon then click to select the other keywords from the list of keywords, and this process takes time. It would be more effective to show users the list of keywords in the interface directly so that they can drag and drop into the text field of each respective keyword category. In addition, by default our prototype system automatically inputs the top keywords into the text field of each category while some of the keywords might not be the acceptable ones. This would take users time to edit and/or remove. Therefore, it would be better to directly show users the list of the keywords in the interface where users can drag and drop in the text field of the respective keyword category. However, not all the keyword candidates should be shown in the first place. For instance, from the above results, we found that if we suggested 5 names to the Who and Org. fields, we will get one acceptable name with the coverage rate of 99% and 49%, respectively. And, for the statistic keywords, if we suggest 5 keywords, we could get 3 acceptable keywords. We also found that when we suggest 29 keywords, we will have 8 acceptable keywords with the coverage rate of 100%. However it is not practical to show all of these keywords. In this case, it is best to show the top 5 suggested keywords. To display other keywords, users just move the mouse pointer to the right or to the left at the end of the suggested keywords zone and other less ranked and high ranked keywords would appear, respectively. Figure 11 shows our proposed interface for the annotation based on our results.

(iii) The information extraction part also takes a great amount of time as it involves lots of natural language phases. Better time gain could be achieved if we were able to perform this task offline.

4.4.2. Discussion

There are a number of issues that the current prototype system does not focus on and they are worth addressing.

(i) We do not concentrate on distinguishing between photos that are taken during subevents which occur within the same time and location, even if they are visually different. Therefore, in our case, for different photos taken on the same date and place, even they are visually different, the same relevant files will be generated. Thus the same candidate keywords will be suggested. However, since we generate a lot of keywords, users can select among the proposed keywords to suit each of the photos in the subevent accordingly. We believe that this is a powerful solution and will make it easier for users to distinguish and recall the events that happen on the same date with automated keywords. Additionally, there are already a number of research efforts in these problem areas such as Naaman et al. [24] that propose algorithms to discover subevents (like a birthday party). Furthermore, using observation and conversation with subjects has allowed us to learn that often subjects do not know which keywords they will eventually attribute to photos. Our system helps users with this task by not only suggesting keywords to associate with photo but also helping them to recall other relevant keywords. Unfortunately, users tend to pick keywords from our suggested terms instead of generating the best new keywords for a given photo.

(ii) Privacy is also a concern. A Google Desktop Search, for example, merely indexes all the files that it has access to. However, should a user with administrative rights install and run GDS within a multiuser environment, the program indexes and searches all files regardless of their owner. We aim to address this problem in our future work.

(iii) To build a faster prototype, we need to rely partially on a number of open source APIs, and we have tried to select the best ones as our performance depends on them.

5. OTHER FEATURES

Besides the annotation engine, we have also built the searching and browsing engines.

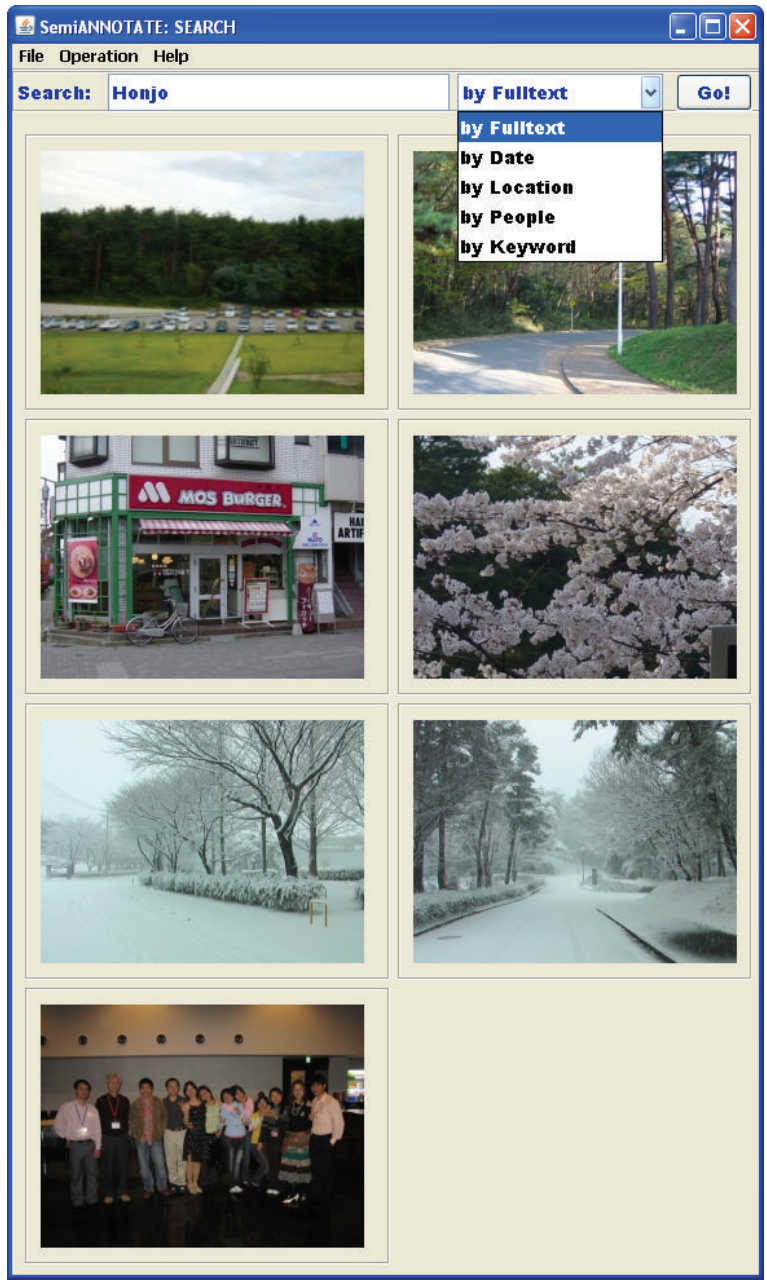


Figure 12: Search engine.

5.1. Searching

We provide five kinds of search, namely, by people's name, date, location, keyword, and full-text. We perform query of each category against our eXist XML native database by using XQuery and XPath. By default, a full-text search is performed to match the input keywords against the entire metadata. Figure 12 shows our proposed searching interface.

5.2. Browsing

We have also built an experimental browsing system based on the episodic metadata that we get from our annotation engine. We believe that we are offering a flexible browsing interface that is different from conventional ones.

In our case, we divide the browsing categories into four: time, location, people's name, and keyword. Users combine the metadata of these different categories to refine the photo sets until they reach the photo that they would like to see. They can go deeper or return backwards. With our interface, navigation becomes much easier for users. The interface gives hints at every stage of the browsing process by showing the possible metadata candidates of each category. Thus users have an easier overall browsing experience. Figure 13 depicts our proposed browsing interface.

Figure 13: Browse engine.

6. RELATED WORK

This section provides the background for the research described in this paper and the context within which the work is situated. The image annotation techniques that have been investigated thus far can be categorized into three major types: manual, semiautomatic, and automatic.

6.1. *Manual annotation (with UI enhancement)*

There are many image management tools (both commercial and research prototypes) that offer the manual annotation capability. What follow are descriptions of several selective systems that represent the essential functionalities of the currently available tools.

Adobe Photoshop Album [4] allows users to define customized keyword tags for people, places, or events and drag them onto photos so that they can be searched later using these tags. Tags can be separated into categories and subcategories for convenient annotation and dynamic organization of photos. Although the annotation system is limited, it is still more effective than the folder-based approach. On the other hand, the annotation process in Google's Picasa [25] and ACDSee [26] is still very time consuming. Users are required to input keywords manually. They only improve the look-and-feel of the GUI of their tools.

One research prototype, PhotoFinder [27], features a drag-and-drop technique that enables users to drag terms (such as person's name) and place them on an image. PhotoFinder associates annotation with coordinates in each photo that later allow for search queries such as "Paul next to Jennifer." On the other end of the spectrum is Caliph, which is part of the Caliph & Emir project [28]. Caliph is a semantic annotation tool designed to help users define semantic objects to be associated with their photos that can later be reused. Caliph can also perform efficient retrieval via the Emir tool.

Collectively, the two obvious burdens of these techniques are that they are time intensive and tedious. In addition, users need to pay great attention during the annotation process in order for it to be effective.

6.2. *Semiautomatic annotation (including collaborative annotation)*

Semiautomatic techniques suggest some pieces of information to users in regards to arranging and clustering photos rather than having the users input everything themselves.

Wenyin et al. proposed the MiAlbum [29] system, which uses feedback to progressively improve annotation in the search process. When a user submits a keyword query, three kinds of results will be generated on the screen: images relevant to the keyword, images that are visually similar to the relevant images, and randomly selected images. A user judges the resulting images using a thumb-up icon. If the user is satisfied, the search keyword will be attributed to that image. The overall quality of the annotations is improved with the extended use of such a system.

The MMM framework [30] allows camera phone users to annotate their photo immediately at the location where they captured the image. This system first displays time and location information and then generates other information from prepopulated lists that others have previously populated with their data through collaborative sharing of tags. A similar strategy is also employed in online photo management systems such as Yahoo!, ZoneTag [31].

Naaman et al. [32] have presented a system that suggests identities inside a photo using the co-occurrence and reoccurrence patterns. The work assumes that accurate location information is available to the photo in addition to date/time information. The method relies on the identities that have previously been associated to the other photos in the collection.

Photocopain, created by Tuffield et al. [5], aims to take advantage of available information such as EXIF metadata, calendar data, community tags, and GPS. However, there is more focus on content analysis than context, and only a few kinds of contextual information are taken into consideration. The work is still in an early stage.

There are many other interesting approaches in this category, but we focus here on those that are closely related to our work. Other methods, such as the SmartAlbum system, assume that each photo comes with voice annotation, and the work analyzes speech signal using speech recognition methods [33]. Girgensohn et al. [34] use face recognition techniques to facilitate the annotation of people appearing inside the photos. The major concerns with these types of systems stem from the fact that most of them only target one aspect of the semantic information, thus creating a lack of scalability for practical implementation.

6.3. Automatic annotation

Many of today's image search engines, such as Google Image Search [35], use surrounding text as a way to generate metadata for the vast number of images on the web. In the web image domain there are an increasing number of investigative systems. One such recent system, AnnoSearch [36], does the annotation first by using an accurate initiative keyword obtained from file names or surrounding text in order to search for other web images. Then, the resultant images are compared and clustered visually and semantically. Li and Wang have proposed an automatic linguistic indexing of pictures or real-time (ALIPR) [37]. This system is an automatic image annotation system that learns from the training dataset and users and is able to achieve significant results in both time and accuracy. Zhou et al. have created an interactive approach for image annotation by incorporating keyword correlations and region matching [38]. However, the results could still be improved upon as well.

Aria [39] enables users to annotate their photos while composing emails. It automatically adds annotation to relevant photos in a collection as the email is being written. This is done using the information from a common sense database [40].

In conclusion, the systems currently in use are a part of a positive trend, and tools of this kind which do not require user intervention are very much needed. However, these systems are still in need of work, as the annotations are most often vague and inaccurate.

6.4. Summary

Despite the diversity of efforts made in the previously mentioned work, the main challenge in generating annotation that represents an individual's interpretation of their photos remains unsolved. So goes the saying, "A picture is worth a thousand words." In an ideal world where a perfect object/face recognition algorithm exists, a computer would still not be able to mimic an individual's perception about a photo without considering its context. The Photocopain system nearly succeeds in integrating contextual information with annotation. However, a perfect system will need to go one level deeper and pay close attention to integrating all available information *to* and *from* users in their ambient environment. The systems presented here are trying to achieve this goal.

7. CONCLUSIONS

A computerized system that accurately suggests annotations or keywords to its users can be very useful. If a user is too busy to create his own keywords, he or she can simply select proposed relevant keywords from a computerized list and add a few more of his/her own. In this paper, we propose a novel and practical paradigm for responding to this type of user demand. We generate contextual keywords for photos from readily available *public and personal sources*, modeling the belief that a user is generally the best authority for describing his or her own photographs and that these user descriptions can usually help generate an accurate interpretation of most photos. Our experiments were conducted on 10 subjects with 313 photographs and the results have proven our theories correct. Our proposed approach contributes to this outcome in three notable ways.

(1) It helps reduce semantic gaps. This is because some parts of keywords are their own keywords (personal information) and the remaining parts are those that they are familiar with, obtained from the news, encyclopedias and other sources (public information). Additionally, we introduce the use of named entities to capture the exact meaning of keywords.

(2) It semiautomates the annotation task rather than working manually. This system also helps the user to recall events with suggested keywords.

(3) It provides a practical implementation framework. This approach is straightforward and is entirely unsupervised. No supervised learning is required to train a prediction of metadata for annotation.

Additionally, we are working to extract more categories of metadata, such as objects (animate and inanimate), events, feeling, actions, numbers, and their relationships. Understanding the relationships between these keywords of different categories will enhance our existing metadata. Furthermore, the methods described in our "Related Work" section can be complementary to this work. Finally, the methodology presented in this paper can easily be extended to the other personal media such as video, text, and audio residing on one's computer.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their constructive comments that helped us to improve the manuscript. It is noted that parts of this paper were presented at the IEEE Int. Conf. on Multimedia and Expo, Beijing, 2007 [41].

REFERENCES

[1] M. Naaman, A. Paepcke, and H. Garcia-Molina, "From where to what: metadata sharing for digital photographs with geographic coordinates," in *Proceedings of the 11th International Conference on Cooperative Information Systems (CoopIS '03)*, pp. 196–217, Catania, Sicily, Italy, November 2003.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[3] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*, pp. 319–326, Vienna, Austria, April 2004.

[4] "Google Image Labeler," 2008, http://images.google.com/imagelabeler/.

[5] M. Tuffield, S. Harris, D. P. Dupplaw, et al., "Image annotation with photocopain," in *Proceedings of the 1st International Workshop on Semantic Web Annotations for Multimedia (SWAMM '06)*, Edinburgh, UK, May, 2006.

[6] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, p. 60, 2008.

[7] "Exif and related resources," http://www.exif.org/.

[8] "Geonames," 2008, http://www.geonames.org/.

[9] C.-T. Lu, M. Shukla, S. H. Subramanya, and Y. Wu, "Performance evaluation of desktop search engines," in *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI '07)*, pp. 110–115, IEEE Systems, Man, and Cybernetics Society, Las Vegas, Nev, USA, August 2007.

[10] "Google Desktop Search," 2008, http://desktop.google.com/.

[11] B. Turnbull, B. Blundell, and J. Slay, "Google desktop as a source of digital evidence," *International Journal of Digital Evidence*, vol. 5, no. 1, pp. 1–12, 2006.

[12] "Google Desktop Search—Search Across Computer," 2008, http://desktop.google.com/features.html.

[13] "Mainichi News," 2008, http://mdn.mainichi-msn.co.jp/.

[14] "Asahi English News," 2008, http://www.asahi.com/english/.

[15] "Wikipedia," 2008, http://en.wikipedia.org/.

[16] "HTTrack," 2008, http://www.httrack.com/.

[17] "Google Desktop Search Java API," 2008, http://sourceforge.net/projects/gdapi.

[18] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: a framework and graphical development environment for robust NLP tools and applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL '02)*, Philadelphia, Pa, USA, July 2002.

[19] "Apache Lucene," 2008, http://lucene.apache.org/.

[20] "Dublin Core Matadata Initiative," 2008, http://dublincore.org/.

[21] "MPEG-7 Overview," http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[22] P. Salembier and J. Smith, *Overview of Multimedia Description Schemes and Schema Tools*, Addison-Wesley, Reading, Mass, USA, 2001.

[23] "eXist XML Database," 2008, http://exist.sourceforge.net/.

[24] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina, "Automatic organization for digital photographs with geographic coordinates," in *Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries (JCDL '04)*, pp. 53–62, Tucson, Ariz, USA, June 2004.

[25] "Picasa," 2008, http://www.picasa.com/.

[26] "Acdsee pro.," 2008, http://www.acdsystems.com/.

[27] H. Kang and B. Shneiderman, "Visualization methods for personal photo collections: browsing and searching in the PhotoFinder," in *Proceedings of the International Conference on Multimedia and Expo (ICME '00)*, vol. 3, pp. 1539–1542, New York, NY, USA, July 2000.

[28] M. Lux, J. Becker, and H. Krottmaier, "Semantic annotation and retrieval of digital photos," in *Proceedings of the 15th International Conference on Advanced Information Systems Engineering (CAiSE '03)*, Klagenfurt, Austria, June 2003.

[29] L. Wenyin, Y. Sun, and H. Zhang, "MiAlbum—a system for home photo management using the semi-automatic image annotation approach," in *Proceedings of the 8th ACM International Conference on Multimedia*, pp. 479–480, ACM Press, Los Angeles, Calif, USA, October 2000.

[30] R. Sarvas, E. Herrarte, A. Wilhelm, and M. Davis, "Metadata creation system for mobile images," in *Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services (MobiSys '04)*, pp. 36–48, ACM Press, Boston, Mass, USA, June 2004.

[31] "Yahoo! Research Berkeley, ZoneTag," http://zonetag.research.yahoo.com/.

[32] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke, "Leveraging context to resolve identity in photo albums," in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)*, pp. 178–187, ACM Press, Denver, Colo, USA, June 2005.

[33] T. Tan, J. Chen, P. Mulhem, and M. Kankanhalli, "SmartAlbum: a multi-modal photo annotation system," in *Proceedings of the 10th ACM International Conference on Multimedia*, pp. 87–88, ACM Press, Juan-les-Pins, France, December 2002.

[34] A. Girgensohn, J. Adcock, and L. Wilcox, "Leveraging face recognition technology to find and organize photos," in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '04)*, pp. 99–106, ACM Press, New York, NY, USA, October 2004.

[35] "Google Image Search," 2008, http://images.google.com/.

[36] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "AnnoSearch: image auto-annotation by search," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 1483–1490, IEEE Computer Society, New York, NY, USA, June 2006.

[37] J. Li and J. Z. Wang, "Real-time computerized annotation of pictures," in *Proceedings of the 14th Annual ACM International Conference on Multimedia (MM '06)*, pp. 911–920, ACM Press, Santa Barbara, Calif, USA, October 2006.

[38] X. Zhou, M. Wang, Q. Zhang, J. Zhang, and B. Shi, "Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR '07)*, pp. 25–32, ACM Press, Amsterdam, The Netherlands, July 2007.

[39] H. Lieberman, E. Rosenzweig, and P. Singh, "Aria: an agent for annotating and retrieving images," *Computer*, vol. 34, no. 7, pp. 57–62, 2001.

[40] P. Singh, "The public acquisition of commonsense knowledge," 2001, citeseer.ist.psu.edu/singh02public.html.

[41] S. Sarin, T. Nagahashi, M. Tadashi, and W. Kameyama, "Exploiting Users' Personal and Public Information for Personal Photo Annotation," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2007)*, pp. 564–567, Beijing, China, July 2007.