



This article was originally published in a journal published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues that you know, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>

Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition [☆]

Bongwon Suh ^{*}, Benjamin B. Bederson

*Human Computer Interaction Laboratory, Computer Science Department, University of Maryland,
A.V. Williams Building, College Park, MD 20742, USA*

Received 12 October 2006; received in revised form 22 January 2007; accepted 20 February 2007
Available online 2 March 2007

Abstract

Managing a large number of digital photos is a challenging task for casual users. Personal photos often don't have rich metadata, or additional information associated with them. However, available metadata can play a crucial role in managing photos. Labeling the semantic content of photos (i.e., annotating them), can increase the amount of metadata and facilitate efficient management. However, manual annotation is tedious and labor intensive while automatic metadata extraction techniques often generate inaccurate and irrelevant results. This paper describes a semi-automatic annotation strategy that takes advantage of human and computer strengths. The semi-automatic approach enables users to efficiently update automatically obtained metadata interactively and incrementally. Even though automatically identified metadata are compromised with inaccurate recognition errors, the process of correcting inaccurate information can be faster and easier than manually adding new metadata from scratch. In this paper, we introduce two photo clustering algorithms for generating meaningful photo groups: (1) Hierarchical event clustering; and (2) Clothing based person recognition, which assumes that people who wear similar clothing and appear in photos taken in one day are very likely to be the same person. To explore our semi-automatic strategies, we designed and implemented a prototype called SAPHARI (Semi-Automatic PHoto Annotation and Recognition Interface). The prototype provides an annotation framework which focuses on making bulk annotations on automatically identified photo groups. The prototype automatically creates photo clusters based on events, people, and file metadata so that users can easily bulk annotation photos. We performed a series of user studies to investigate the effectiveness and usability of the semi-automatic annotation techniques when applied to personal photo collections. The results show that users were able to make annotations significantly faster with event clustering using SAPHARI. We also found that users clearly preferred the semi-automatic approaches.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Human–computer interaction; Annotation; Semi-automatic annotation; Event identification; Clothing based person recognition; Zoomable user interface; Face detection; Digital photograph; Image management systems

1. Introduction

As the use of digital cameras becomes more popular, more images are created and stored in personal computers and shared over the Internet, and managing these photos

becomes proportionally more difficult. One reason for the difficulty can be attributed to the fact that photos in a personal collection often do not have rich metadata that describes their content. Photos taken by most digital cameras often include only basic low-level metadata (JEITA, 2002) such as focal length, time, and resolution.

On the other hand, once available, high-level metadata associated with photos can play a crucial role in managing photos. It enables users to search, browse, and organize photos very efficiently. For example, it is obvious that high-level metadata such as “William’s birthday party” is more useful and appropriate to users than basic

[☆] This research was performed when the first author was a doctoral student at the University of Maryland, College Park, MD. The first author is now with Palo Alto Research Center, Palo Alto, CA.

^{*} Corresponding author. Tel.: +1 650 812 4806; fax: +1 650 812 4258.

E-mail addresses: bongwon@acm.org (B. Suh), bederson@cs.umd.edu (B.B. Bederson).

information such as an arbitrary filename, “IMG_1997.JPG”.

Hence, many researchers have worked on various techniques to extract relevant metadata directly from images including: event identification, object identification, face detection/recognition, content-based categorization and so on (Cooper et al., 2003; Naaman et al., 2004; Riya.com, 2005; Yang et al., 2002; Yoshitaka and Ichikawa, 1999; Zhao et al., 2003). However, automatic metadata extraction is often inaccurate. In addition, the obtained metadata may be too general to satisfy the needs of every individual user. Each user needs different types of metadata according to his/her own interest. Furthermore, there are numerous cases where it is not possible at all to automatically obtain metadata without the intervention of humans. Identifying an event, such as a birthday party or its location would be particularly hard.

On the other hand, manual annotation, defined as a process which involves labeling the semantic content of images (or objects in images) with a set of keywords or semantic information, can directly associate metadata with photos. However, this is labor intensive and tedious. One approach that is gaining interest is to harness the skills of anonymous volunteers to manually annotate photos over the web. This has been done to great success (Von Ahn et al., 2006). However, that kind of approach will not work for personal photos that require private knowledge of the related people and events in addition to being able to recognize features of the photo. A valuable acceleration technique is bulk annotation, in which a group of photos is selected and then the same label is applied to every photo with one action (Kang et al., *accepted for publication*). However, this approach requires that related photos that should be annotated with the same label need to be presented to the user near each other to enable efficient selection. This is somewhat of a chicken and egg problem as this visual clustering requires some metadata to begin with.

We propose instead, a semi-automatic photo annotation framework that focuses on facilitating bulk annotation on photo groups that have been automatically identified. Semi-automatic annotation combines manual and automatic metadata extraction. The semi-automatic approach allows users to update automatically obtained metadata interactively and incrementally. Even though automatically identified metadata are compromised with inaccurate recognition errors, the process of correcting inaccurate information can be faster and easier than manually adding new metadata from scratch. The process of correcting errors can be even faster if it is assumed that users often focus on the important errors and disregard the less important ones.

To explore semi-automatic strategies, we designed and implemented a prototype called SAPHARI (Semi-Automatic PHoto Annotation and Recognition Interface). The prototype provides an annotation framework that allows users to make bulk annotations on automatically identified photo groups. The prototype automatically creates photo

clusters based on events, people, and image file directory structures so that users can easily annotate useful information such as events, location, and people. We focused on people and events because of Rodden and Wood’s analysis of casual personal photo patterns of use (Rodden and Wood, 2003). They found that “event” and “person” are the most frequently used metadata as well as chronological order of photos.

In this paper, we introduce two strategies to identify meaningful photo groups to assist users in efficiently annotating those types of metadata. Our first method is to cluster photos by their timestamps so that each group of photos can correspond to a meaningful “event”. The proposed hierarchical event clustering identifies multiple event levels in personal photo collections and allows users to choose the right event granularity. The second approach is to use a clothing-based person recognition technique to group similar-looking people together. The clothing based person model assumes that people who wear similar clothing and appear in photos taken the same day are very likely to be the same person.

We also performed a series of user studies to investigate the usefulness of the semi-automatic annotation techniques when applied to personal photo collections. These studies showed that users were able to annotate significantly faster with SAPHARI when annotating event information. Users also clearly preferred the proposed semi-automatic approaches.

2. Related work

2.1. Image management systems

We became motivated to work on this problem through PhotoMesa (Bederson, 2001), a tool we built previously. It allows users to view a large set of image on one screen using a Zoomable User Interface, made it very clear to us that acquiring good metadata was crucial to the success of people using their digital photos. PhotoMesa allows users to view multiple directories of images with zooming and panning navigation. Users can zoom in to see the detail image and zoom out to view the overview of images. PhotoMesa is designed to provide an intuitive and efficient image browsing environment. PhotoMesa displays multiple groups of images on the screen at once by laying them out in a 2D zoomable space with a Quantum Strip Treemap algorithm (Bederson, 2001).

There are a number of other research prototypes that also address metadata issues. FotoFile is a multimedia organization and retrieval prototype that allows users to browse photos grouped in albums (Kuchinsky et al., 1999). When an album is selected, images in the album are laid out on the screen for viewing and editing. Through informal user studies, Kuchinsky et al. (1999) found that: (1) users did not want to spend a lot of time organizing their photos with keyboard annotations; and that (2) they wanted to browse through photos, not just perform direct

search activities. Kuchinsky et al. also noted that people like to tell stories with photos.

PhotoFinder (Kang and Shneiderman, 2000) offers a drag-and-drop technique known as direct annotation to enable personal names to be placed on a photo or a group of photos. Users can drag keywords (usually person's name) onto any place on the photos to save typing.

In addition to research prototypes, there are a vast number of commercial products available. Some notable systems are as follows.

Adobe PhotoShop Album (Adobe, 2006) is a product designed to gather all photos in the users' computer and let users see those photos in one convenient place, organized by date or any chosen subject. The system has a timeline showing the distribution of photos over time. The timeline has two sliding knobs and users can filter photos by the date they were taken. Adobe PhotoShop Album incorporates a keyword tag. Users can create customized keyword tags that represent special people, places, or events, and drag them onto photos so that pictures can be found by subject later. These improvements help users make annotations efficiently in comparison with the manual annotation strategy where users are required to type in keywords.

Picasa (2006) also uses a folder concept. When first launched, Picasa searches the entire storage system on a computer and created an album per folder. Users are allowed to organize folders by merging and renaming them. Photos are laid out on the screen by the albums and users can use scroll bars to navigate them. On the left panel of the screen, the list of albums is arranged by timeline. Picasa also supports image editing functions as well as sending pictures quickly by using e-mail clients. While users are allowed to add any keyword to photos in Picasa, the annotation process is very time-consuming. When users are adding keywords, a separate window form is provided, and users are required to type keywords manually. Furthermore, Picasa does not support a list of existing keywords, which makes consistent annotation difficult.

Tagging is also available for an online photo album, Flickr.com (2005). In Flickr, tags can be shared by other users and public photos can be searched by them. However, it is still a burden for users to add tags on a large number of images.

There are many, many other photo management tools as well, but the ones described here capture the essential features commonly available.

2.2. Semi-automatic annotation interface

Others have explored the idea of semi-automatic annotation before. Our approach, however, visually groups similar photos in a single screen zoomable user interface, and includes two novel algorithms for creating those groups.

Wenyin et al. (2001) introduced a semi-automatic approach to make annotations on digital photos. Their

progressive annotation process is embedded in the course of integrated keyword-based and content-based image retrieval. When a user submits a keyword query, the system retrieves and arranges the search results on the screen. The search results include: (1) images that are relevant to the search keyword; (2) visually similar images to the relevant images; and/or (3) a set of randomly selected images. When an image receives positive feedback from users (by clicking the thumb-up icon), the search keywords are automatically added to the images so that the images can be retrieved by keyword-based image retrieval in the future. The coverage and quality of image annotation is improved progressively as the cycle of search and feedback increases. Wenyin et al. (2001) reported that the semi-automatic image annotation strategy is better than manual annotation methods in terms of efficiency, and is better than automatic annotation techniques in terms of accuracy. However, the authors describe the MiAlbum (Wenyin et al., 2001) user interface as needing enhancement. The simple thumbs-up/down metaphor was not enough for users to understand the built-in underlying automatic algorithm. They also report a problem in the discoverability of relevance feedback.

Girgensohn et al. (2004) created a photo management prototype which employs a semi-automatic interface to facilitate labeling people in photos. They used a face detector to automatically extract faces from photos while using a less accurate face recognizer to sort faces by their similarity. The sorted faces are presented as candidates for annotation targets. Users are allowed to drag faces onto name labels to make annotations. Their simulated user study showed that on average 60% of faces could be assigned successfully with three or four steps. While the semi-automatic approach of Girgensohn et al. (2004) shows potential, there are scalability and usability issues. As the number of faces in the system increases, users are required to frequently scroll. Furthermore, as the number of people increases, the face recognition accuracy decreases significantly and it makes bulk annotation harder. In addition, this approach depends entirely on face recognition software which is notoriously poor for general photos. Even with state-of-the-art systems, face recognition accuracy for outdoor photos is around 50% (Phillips et al., 2003).

Naaman et al. (2005) introduced a co-occurrence based annotation mechanism which relies on the availability of accurate location information. Their system suggests likely identity labels for photographs by leveraging automatically-available context such as the time and location where the photos were taken. As the user annotates some of the identities of people in their collection, the system utilizes the patterns of re-occurrence and co-occurrence of different people in different locations and events. The system uses these patterns to generate label suggestions for people that were not yet annotated. They reported that the automatic suggestions can accelerate the process of manual annotation.

Tuffield et al. (2006) developed a photo annotation system, Photocopain, aiming to simplify the photo annotation

task by taking advantage of available context information such as EXIF metadata (JEITA, 2002), GPS, calendar data, and community tags. While Photocopain showed a great potential for classifying photos, more research is required on how to integrate the proposed techniques with manual annotation.

2.3. Event identification

There are a number of approaches to extracting event information from personal photo collections. Cooper et al. (2003) introduced an approach that clusters photos by combining temporal information and image content. While their experimental results show approximately 85% accuracy, it was not significantly better in comparison with other algorithms such as Leung et al. (2000) and Kohonen (1989). Also, using the image content did not make significant contribution to detect events.

Girgensohn et al. (2003) presented a prototype photo manager based on Cooper et al. (2003)'s event detection algorithm. In Girgensohn et al.'s prototype, photographs are presented to users in event-groups which have been automatically identified.

Platt et al. (2002) use an adaptive local threshold method to detect event boundaries. Platt et al.'s algorithm (2002) compares a time interval to its local average interval. If a temporal gap between two adjacent photos is considerably larger than its weighted local average, the algorithm determines that the gap represents an event boundary. Unlike the algorithm in Cooper et al. (2003), this algorithm requires additional parameters including a threshold for sensitivity and a window size. These parameters should be empirically chosen and can be subjective. Cooper et al. (2003) reported that the accuracy of this algorithm was not very good compared to other clustering algorithms.

Scale-space analysis (Leung et al., 2000) is a technique for accessing structure at multiple scales in a data set. It assigns a Gaussian kernel per data point to form a Gaussian mixture. The result mixture is used to form clusters by finding points where its second derivative value is zero (peak point). By using varying standard deviation, it allows to construct hierarchical segmentation.

Loui and Savakis (2000) use the *K*-means algorithm combined with content-based post-processing for automatic albuming of photographs. They checked the color similarity of images at event boundaries to verify that the images indeed differ.

Graham et al. (2002) use time information for creating event hierarchies for personal photo collection. Based on Platt et al. (2002), they create initial clusters. Then, they build an event hierarchy based on the initial clusters. For each cluster, a summarization photograph is selected to represent the event. With this clustering and summarization technique, they built a prototype, "Hierarchical Browser" and performed a user study. They found that users completed given tasks better with the hierarchical browser.

They also showed that the summarization technique significantly reduced users' browsing completion time.

Gargi (2003) presented an analysis of consumer media capture behavior based on timestamp metadata. Gargi (2003) reported bursty behavior of personal photo collections. Date sets used in Gargi (2003) shows that photos are taken approximately one day out of ten. However, on the day that photos are taken, users take about twelve photos at once.

Naaman et al. (2004) uses the time and location information embedded in digital photographs to automatically organize a personal photo collection. They designed a prototype, PhotoCompas, which produces browseable location and event hierarchies for the collection. The hierarchies are created using algorithms that interleave time and location to produce an organization that mimics the way people think about their photo collections. Their case studies report that the photo grouping is meaningful and useful for the collection owners.

As shown, there has been significant research on automatic event identification. Nevertheless, automatically identified events are not perfect and require users' modification. Most of the above approaches did not consider users' feedback on event boundaries. Once event boundaries are set by an algorithm, the presented interfaces are not adaptable and the algorithms do not allow further interaction with users.

3. Metadata and annotation

Annotation is a process of labeling the semantic content of photos (or objects in photos) with a set of keywords or semantic information. Annotated information is very important for managing photos since the information plays a crucial role for organizing, browsing, and searching photos. This association between keywords and the contents of photos can be achieved in a number of ways. They can be roughly categorized into: (1) direct importing; (2) image analysis; (3) extraction from image context; and (4) manual annotation as follows.

Direct importing: Some types of information can be directly obtained from image files or photographic devices. File name, file size, date and EXIF (JEITA, 2002) information can be easily acquired. However, this type of metadata often does not have much value for users, especially for casual users who want to manage their own personal photos. For example, an image file name "IMG_2345.jpg" is not very useful.

Image analysis: Some image features such as colors, texture, and blobs can be extracted by analyzing images. For example, QBIC (Faloutsos et al., 1994) enables users to specify search conditions based on low level visual features such as color and blob location. In QBIC, a user can issue queries such as "find images which have red objects in the center". However, these types of low level metadata are often irrelevant in many cases. For personal photos, higher

level metadata, such as location, event, or person in photos would be more relevant and interesting to users.

Extraction from image context: Context of images often contains useful information about the images. Shen et al. (2000) used the textual context of web pages to extract descriptive information of images on the same pages. This type of approach is potentially useful when applied to images with captions. While this approach can provide rich metadata about images, its application is limited and not applicable for general photographs.

Manual annotation: Users can explicitly decide which information should be added on a specific photo. It is obvious that the actual users, who are annotation makers and metadata consumers at the same time, can function as the most reliable source of accurate and relevant metadata for photos. But, it is well known that most users do not want to spend much time creating and annotating metadata. Kang and Shneiderman (2000) developed a direct annotation method that focuses on labeling names of people in photos. While it saves users typing work, users still have to perform drag and drop many times.

Automatic metadata extraction, including direct importing, image analysis, and extraction from image context is usually fast (compared to manual annotation) but often generates inaccurate and irrelevant results. Furthermore, there are numerous cases where it is impossible to extract metadata from images without the intervention of humans.

The basic idea of semi-automatic annotation is to incorporate users' feedback into metadata that was automatically extracted. When the metadata has reasonable accuracy; the amount of inaccurate information would be considerably less than that of accurate information, and manually fixing inaccurate metadata can be faster and easier than adding new information. Users' interactive and incremental correction (e.g., relevance feedback) also can be used as an input for automatic metadata recognition systems to increase recognition accuracy. As users keep using the system, the overall metadata accuracy, as well as the quantity of metadata, can increase since more reliable metadata are added by the users.

3.1. Design principles for semi-automatic annotation interfaces

Any semi-automatic annotation strategy includes frequent interactions between users and recognition systems. Therefore, it is crucial that user interfaces for semi-automatic annotation should guide users to efficiently interact with systems. We have created the following design principles based on our experience building PhotoMesa.

Bulk annotation: Bulk annotation, where multiple images are annotated with a single command action, can accelerate users' annotation performance. Rather than repeatedly selecting images and making annotations one by one, making annotations on selected multiple images can speed up the annotation process. However, the acceleration is achieved only when selecting multiple annotation

targets is easy and efficient enough. If the annotation target selection takes too long, there would be no benefit. A semi-automatic annotation interface, therefore, should be carefully designed to allow users to choose multiple annotation targets efficiently. For example, when images that share common or similar information are located closely together on the screen, they can be a good candidate for bulk annotation. Semantically close items that are laid out together on the screen often facilitate bulk annotation.

Transparent interface: The relationship between automatic annotation and users' relevance feedback should be made very clear to the user. Koenemann and Belkin (1996) observed that users perform better when they understand the underlying algorithms. They showed that increasing the transparency of relevance feedback helped users take advantage of an automatic algorithm. An interface for semi-automatic annotation systems, therefore, should provide clear information about how it processes information.

Users in control: Users should be in control at all times. Automatic labeling should be a suggestion to users, not a mandatory obligation. Users must have the freedom to make annotations as they choose. A semi-automatic annotation interface should not block or interfere users' manual overriding.

Show contextual information: Providing appropriate contextual information can increase annotation performance. Various contextual information can be provided to help users to make annotations. For example, metadata already associated with a photo such as timestamp, tags can be useful for users to make additional annotations. When photos are chronologically ordered, neighboring photos can provide clues about what a specific photo is. In addition, spatial hierarchies in metadata (i.e., Palo Alto is located in northern California) can be utilized. Naaman et al. (2005) showed that temporal and spatial contextual information is useful for managing digital photos. We found that showing context is especially important when designing semi-automatic photo annotation interfaces. Since automatic labeling sometimes contains inaccurate metadata, users frequently have to correct inappropriate automatic labels. In this case, the context can help users to select proper metadata. However, it is not always better to provide all the available information on the screen. A user interface should prioritize the available information and provide a modest number of alternatives.

Incremental and interactive annotation: Users must not be forced to make annotations. A photo annotation interface should allow a user to make annotations at any time. Users also should be allowed to make annotations on important and interesting photos first and to defer annotations on other photos into later times.

Based on the above design principles, we designed and implemented a semi-automatic photo annotation prototype, SAPHARI (Semi-Automatic PHoto Annotation and Recognition Interface). It is designed to help users manage their personal photo collections by using auto-

matic recognition systems. SAPHARI takes advantage of automatic recognition algorithms when laying out photos on the screen. While our earlier work on PhotoMesa (Bederson, 2001) depends on basic metadata such as directory, date, and filename to form photo groups, SAPHARI uses hierarchical event identification (see Section 4) and clothing based person recognition (see Section 5) in addition to the basic file system structures.

4. Event identification and annotation

Temporal information plays an important role when classifying personal photos since personal photos often are taken near each other in time (Gargi, 2003). In other words, when photos are close in time with each other, they have a high probability of sharing common or similar information. Given the temporal locality in a personal photo collection, photos can be prepared in groups according to their timestamps so that they can be easily bulk-annotated.

As stated earlier, events are one of the most important types of metadata for organizing personal photos. There have been a number of efforts to find meaningful event clusters from image collections (Cooper et al., 2003; Gargi, 2003; Kuchinsky et al., 1999; Naaman et al., 2004; Platt et al., 2002). This type of time based event identification is based on the episodic nature of personal photo taking. When there are interesting things around a user with a camera, the user tends to take a relatively large number of photos in a short period of time. Then, there is often a relatively long pause, followed by another burst of activity. Based on this characteristic, event boundaries are identified by detecting relatively long temporal gaps in a photo collection. The challenge is that the lengths of the events and the lengths of the gaps vary, so fixed thresholds for determining event boundaries do not work well.

4.1. Event hierarchy

Events in a personal photo collection tend to form a temporal event hierarchy (Naaman et al., 2004). In other

words, events in a collection can be defined in multiple ways depending on event granularities. An example can help to clarify (Fig. 1). “Summer Camping Trip”, which spans June 13–17th, can contain multiple sub-events such as “Hiking” on the 14th, “Canoeing” on the 15th, and “Santa Cruz” on the 16th. This event hierarchy can provide useful and flexible event units to users. But, existing event identification techniques such as Cooper et al. (2003) and Platt et al. (2002), which focus on a single level of events, cannot capture all the meaningful events in a photo collection.

Event units, once identified, are very useful for making bulk annotations. In this same example, a user can add the label “Summer Camping Trip” to all photos taken between June 13th and June 17th with a single action if the photos are clustered with coarse granularity. Similarly, the user can add a “Hiking” label to photos taken on June 14th with a single action when photos are tightly grouped with finer event granularity.

The event based annotation can be very effective if all identified events match users’ intention. However, automatic event identification does not necessarily correspond to all events that users want to have. We identified a set of operations that users can perform when they find errors in automatically identified event boundaries. Users can modify events by either: (1) splitting an event group, (2) merging two neighboring events, or (3) moving event boundaries. For example, if photographs taken on June 15th (Fig. 1) have to be regrouped, they can be: (1) merged with neighboring event groups (“Hiking” or “Santa Cruz”), (2) split into more detailed events, or (3) changing event borders (when event detection algorithms fail to identify correct event boundaries).

When two events are merged or an event is to be divided, event boundaries on different levels can be also affected. When a user merges events or splits an event, event boundaries on different levels are automatically updated to satisfy the following two conditions: (1) when events are merged at a finer level, the event groups cannot be split at coarser levels; and (2) when events are split at a coarser level, those events cannot be merged in finer levels.

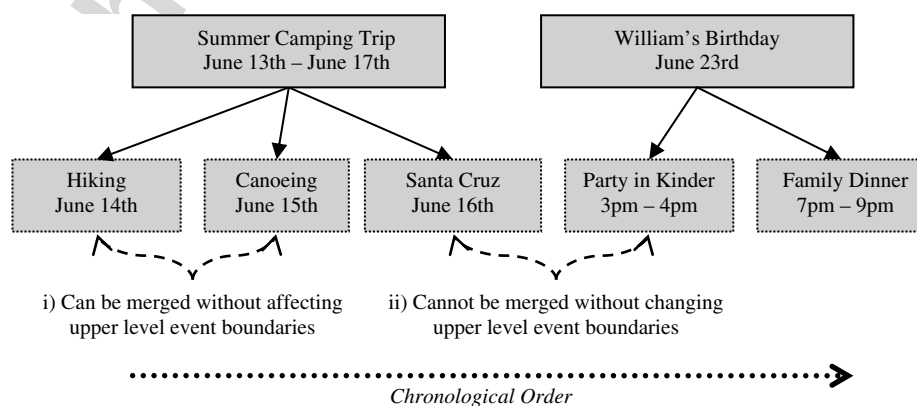


Fig. 1. An example event hierarchy. The units in the upper row are coarsely grouped and consist of units in the lower row, which are tightly grouped. Users are allowed to modify the event grouping. However, merging neighboring events may affect event boundaries at different levels.

This type of event boundary propagation is integrated with photo browsing environment, which will be explained in the next section.

4.2. Event identification algorithm

To identify a hierarchical event structure in a personal photo collection, we developed an interactive and adaptive algorithm (Fig. 2), which supports multiple event levels. The hierarchical event identification algorithm enables more flexible grouping by supporting event boundary modification. The algorithm employs an event boundary detection algorithm based on Platt et al. (2002). However, the proposed algorithm can be used to identify a hierarchical event structure and multiple levels of events integrated with users' knowledge, while the original Platt et al. algorithm focuses only on detecting single level event boundaries.

The algorithm considers event boundaries of different event levels as well as identifying unusually long temporal

gaps between photos. Event detection granularity can be controlled by changing the K and d parameters. In SAPHARI, we use empirically chosen $K = \{25, 200\}$ and $d = \{10, 20\}$.

As shown in Fig. 3, when users choose “Event” layout, SAPHARI lays out all photos in a collection, grouped by identified events. SAPHARI provides two granularity options for event clustering, *Fine* and *Regular*, which users can switch between on demand. As the event grouping granularity changes, SAPHARI immediately changes its grouping and rearranges photos on the screen.

In this event view mode, SAPHARI allows users to browse a photo collection with zooming and panning functions as in PhotoMesa. While browsing, users can opt to fix inaccurate event boundaries if they find the identified events are inadequate. The users' boundary fixing is automatically propagated to different events levels of the event hierarchy. For example, merging two neighboring events in the *Fine* event view automatically aggregates photo groups in the *Regular* event view.

```

BUILD_HIERARCHICAL_EVENT_CLUSTER(images, current event level)
  foreach image in images
    If (currentGroup is null) {
      currentGroup = new EventGroup();
    }

    If (image.eventBoundary[finer granularities].merge is true) {
      // Case 1: Use the group of the previous image
      image.eventboundary[current event level].merge = true;
    } else if (image.eventBoundary[coarser granularities].split is true){

      // Case 2: Create a new group
      image.eventboundary[current event level].split = true;
    } else if (image.eventboundary[current event level] is not defined) {
      // Case 3: Use the algorithm to decide
      // whether to create a new group
      image.eventboundary[current event level].split =
        SPLIT_BEFORE(images, index, current event level)
    } else {
      // Case 4: Keep the previous grouping
      // keep the current image.eventboundary[current level]
    }

    if(image.eventboundary[current event level].split) {
      // Create a new group and add the current image into it
      currentGroup = new EventGroup();
      currentGroup.add(image);
    } else if(image.eventboundary[current event level].merge) {
      // Add the current image into the group of the previous image
      currentGroup.add(image)
    }
  end foreach
END

BOOL SPLIT_BEFORE(images, i as current index, l as current event level)
  T[] = ordered timestamps collected from images
  Kl = K[l]
  dl = d[l]
  If  $\log(T[i] - T[i-1]) \geq K_l + \frac{1}{2d_l + 1} \sum_{j=-d_l}^{d_l} \log(T[i+j] - T[i+j-1])$ 
    return true
  Else
    Return false
END

```

Fig. 2. Pseudo code for building hierarchical event clusters.

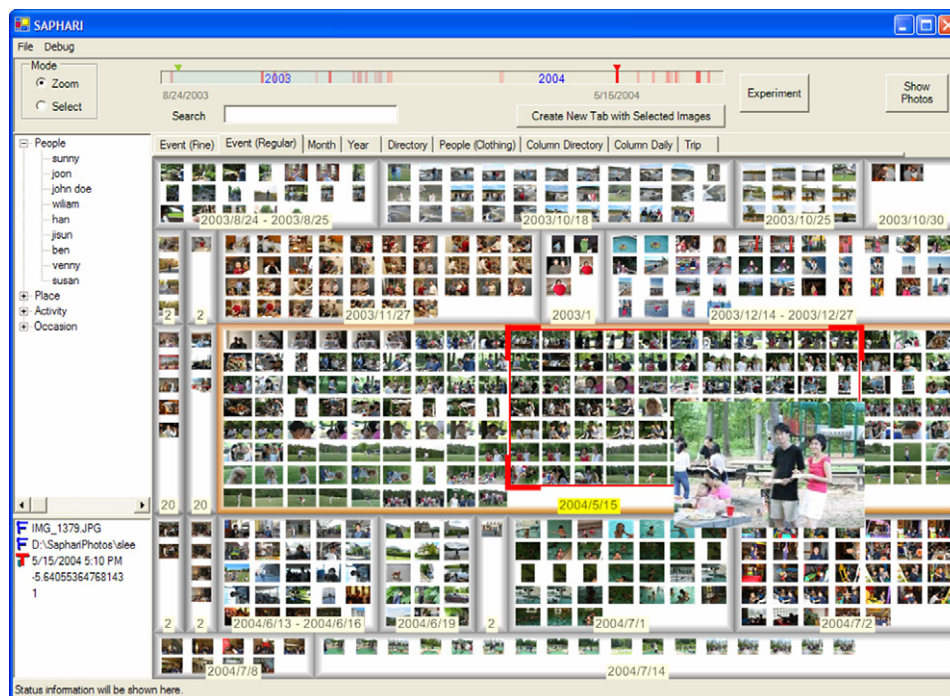


Fig. 3. Photos in a collection are organized in a 2D zoomable space, grouped by the hierarchical event identification algorithm. The identified photo groups are ordered chronologically from the earliest to the latest with groups laid out in rows left to right, and top to bottom. Users can browse a photo collection by zooming and panning. A preview of the image under the mouse cursor is shown when hovering the mouse.

4.3. Annotation strategy

Since bulk annotation is a valuable accelerator for creating metadata, SAPHARI is designed to facilitate users in bulk-annotating event information. SAPHARI arranges photos on the screen by event groups and annotation is performed by drag-and-dropping a text label onto a target photo or a group of photos.

Fig. 4 shows an example annotation procedure, adding “Santa Cruz” to a single photo. Annotation begins when a user selects a text label. In Fig. 4, “Santa Cruz” is selected from the metadata list on the left side of the screen. With the mouse dragged, the cursor is changed into the text of the selected label, “Santa Cruz”. At the same time, the drop target is highlighted to give visual feedback to the user. In Fig. 4, the borders of a photo under the mouse cursor are highlighted with an orange color. As with the stan-

dard drag-and-drop metaphor, the metadata is annotated onto a highlighted photo as the mouse button is released.

With the shift key pressed, SAPHARI chooses all the photos in the event group under the mouse cursor as its drop target instead of a single photo. When a label is dropped on a group, all photos in that group are annotated with the dragging label at once (Bulk annotation). SAPHARI also supports bulk annotation on any arbitrary pre-selected group of photos.

When the automatic event identification fails to detect events accurately (Cooper et al., 2003), SAPHARI allows users to manually override any event boundaries. Fig. 5 shows utility functions that SAPHARI provides as a context menu for supporting the event boundary modification.

5. Clothing based person recognition

It is not surprising that many image browsing prototypes (Adobe, 2006; Kang and Shneiderman, 2000; Shen et al., 2001; Riya.com, 2005) focus on labeling people with metadata such as names since the people in personal photos are extremely important. As discussed in the previous section, grouping is important for efficient bulk annotation. SAPHARI is designed to help users to bulk-annotate people in photos by grouping faces of the same person.

5.1. Face recognition for personal photos

Researchers have had limited success in developing software to accurately recognize faces in images of uncontrolled

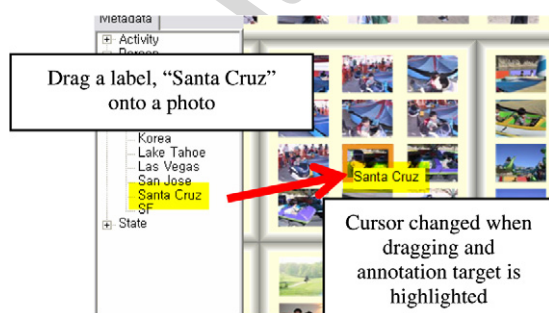


Fig. 4. Annotation by drag-and-drop. Users can drag a text label onto a photo or a group of photos to make annotations.

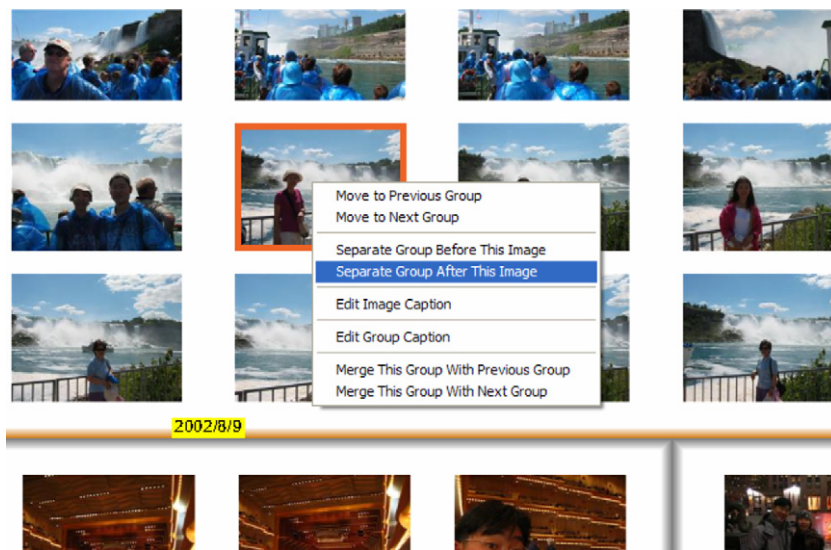


Fig. 5. Fixing automatically identified event boundaries. Context menu provides users with a set of functions that can control event boundaries.

environments commonly found in personal photos. For example, even for the best face recognition systems, the recognition rate for faces captured outdoors was only about 50% at a false rate of 1% (Phillips et al., 2003).

During our preliminary research, we used the HMM face recognizer included in OpenCV (2005) to classify people in photos. Even though the face recognizer produced reasonable results when applied to controlled face sets (taken indoors with controlled lighting and frontal views), the accuracy dramatically decreased when used on personal photos. The face recognizer was very sensitive to lighting condition and tilted faces. People in personal photo collections sometimes are not gazing at the camera, which aggravates the hardship in face recognition. Therefore, we concluded that we cannot solely rely on human face recognition to identify people in personal photos.

As an alternative, we observed an interesting pattern in personal photo collections. People usually do not change their clothing in a given day. Given this condition, clothing information can be used to assist the identification of people.

5.2. Human model

Naturally, clothing can be located on human bodies. However, locating human bodies from a static scene is not an easy task because the shape of the human body is not rigid and requires complicated modeling. While many researchers have worked on detecting human body movement, it is still challenging to reliably detect human body parts from a single static scene (Barron and Kakadiaris, 2001; Rosales and Sclaroff, 2000).

SAPHARI avoids the complexities of human body modeling by using face detection to inform the location of clothing in photographs. Compared to face recognition systems which have low accuracy, face detection systems

(whose goal is to identify the location of faces) have higher accuracy (Yang et al., 2002). Some systems have reached around 90% accuracy for detecting faces in images. We use the Viola–Jones face detector (Lienhart and Maydt, 2002; OpenCV, 2005; Viola and Jones, 2004) to locate faces in a photo.

We developed a clothing model based on the detection of human face location (Fig. 6). Since the face detector also provides the face size, we can base the upper body clothing region is calculated based on the size of the face. As shown in Fig. 6, we construct the human model only with upper body clothing. Theoretically, it would be optimal to use the whole body information. However, as explained earlier, identifying a human body in a static scene is not an easy task and is beyond the scope of this paper. Furthermore, in many personal photos, we found the body to be cropped so that only the upper body is visible.

Once a face is detected, the area directly underneath the face is skipped as we assume that area belongs to a neck. Then, aligning with the center line of the face, we select a rectangular region as clothing. The proposed clothing based person model is built from pixel samples picked from the clothing region. We found that picking samples from the upper central part of the clothing provides more robust results than selecting samples uniformly from all parts of the clothing. Therefore, when picking up pixels in the clothing area, more weight is given along the center line (horizontal sampling weight) and to the upper part of the clothing (vertical sampling weight). The horizontal weight is given by using a normal distribution whose standard deviation is $3/8$ of the face width. The vertical weight is linearly distributed with the topmost row having twice as much weight as the bottom-most row.

In this human model, a four dimensional feature vector $X = (y\text{-distance, red, green, blue})$ is employed to model the clothing, where $y\text{-distance}$ is defined as a relative vertical

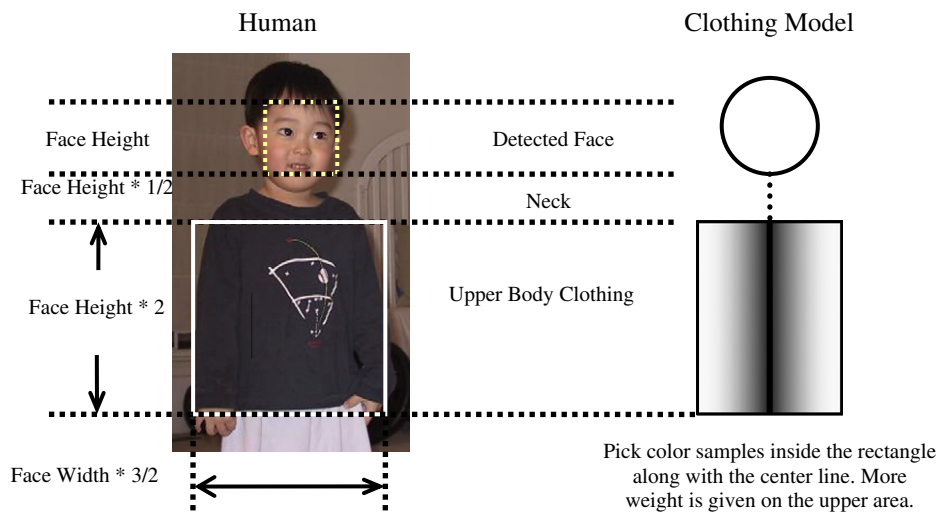


Fig. 6. Locating clothing from detected faces.

position of a sample and red, green, and blue are color value of the sample pixel, respectively. In other words, each selected pixel in the clothing area is transformed into a four dimensional vector. In our model, x -distance from the center line is not included because adding it did not improve the performance of the system while requiring significantly more computation. We also found that adding x -distance makes our model susceptible to tilted faces and bodies.

About 900 samples are collected from the upper body region to form a single model. Based on the collected four dimensional vectors, we estimate a four dimensional probability density function per clothing of a person using the following kernel density estimation formula (Silverman, 1986):

$$f(X) = \frac{1}{n} \sum_{i=1}^n \frac{1}{(2\pi)^{d/2} |H_i|^{d/2}} \exp \left(-\frac{1}{2} (X - X_i)^T H_i^{-1} (X - X_i) \right),$$

where n is a number of samples and d is 4, respectively. H is a diagonal matrix with independent four variances as follows.

$$H = \begin{bmatrix} \sigma_{y\text{-distance}} & 0 & 0 & 0 \\ 0 & \sigma_{red} & 0 & 0 \\ 0 & 0 & \sigma_{green} & 0 \\ 0 & 0 & 0 & \sigma_{blue} \end{bmatrix}$$

As shown in H , we assume that there is no correlation between y -distance and red, green, and blue. For the prototype, we use empirically determined parameters, $\sigma_{y\text{-distance}} = 0.08$ (of the clothing height), $\sigma_{red} = 0.04$, $\sigma_{green} = 0.04$, and $\sigma_{blue} = 0.04$, respectively.

Clothing based person recognition utilizes an estimated four dimensional probability density function (pdf) as a proxy for a person. We can then measure the visual distance between two pieces of clothing using the Bhattachar-

ya distance (Bhattacharya, 1943) which measures the distance between two pdf s (Fig. 7).

SAPHARI classifies the people in photos using these distances. When a measured distance between two models is below an empirically determined threshold, the system classifies the two models to be the same. When the distance between two models is above a threshold, SAPHARI catalogs them as two different people. SAPHARI uses an empirically chosen threshold of 0.4.

When a user chooses the “Clothing” tab in SAPHARI, the user is asked to pick one date among the list of dates when photos was taken. SAPHARI analyzes the faces and clothing of the images from that display, and shows groups of each identified person’s face and clothing (Fig. 8). Each cropped face is normalized (shrunk or enlarged) to have the same size on the screen for easy comparison. As shown in Fig. 8, when hovering the mouse on a face, a preview of the original photo is shown with the face highlighted for context.

5.3. Annotation strategy

Since people are grouped together, users can make bulk annotations on a group of people with a single action by dragging a name label onto a face group. Upon dropping a name label onto a group, all faces in the group under the cursor are annotated with the name (Fig. 9). After a name is assigned to a face group, the group itself is associated with the name.

While better than face recognition, the clothing recognition still generates errors, so the interface allows users to correct them. As shown in Fig. 10, users are allowed to move a face image or a group of face images into another group. Once face images are moved into other groups, the face images update their labeling association. This concept, where a group is associated with a set of metadata, is introduced by Kang (2003) and called “semantic regions”.

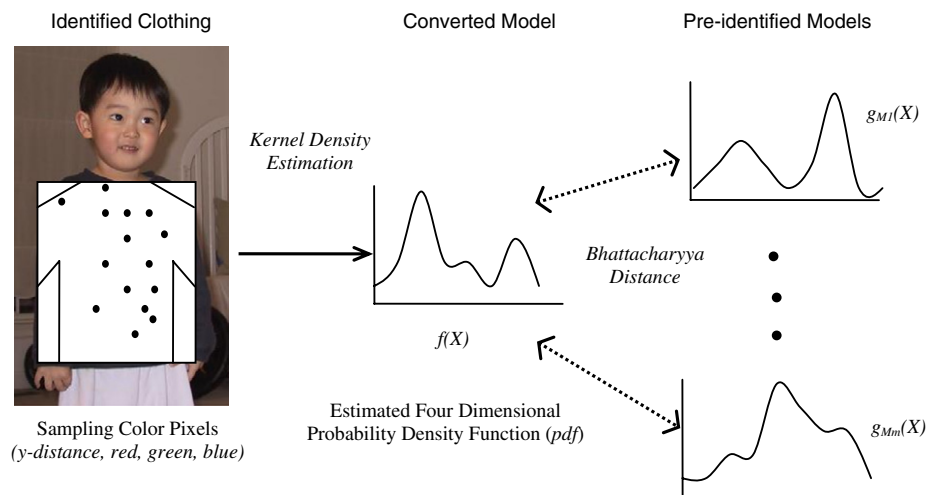


Fig. 7. Comparing clothing models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

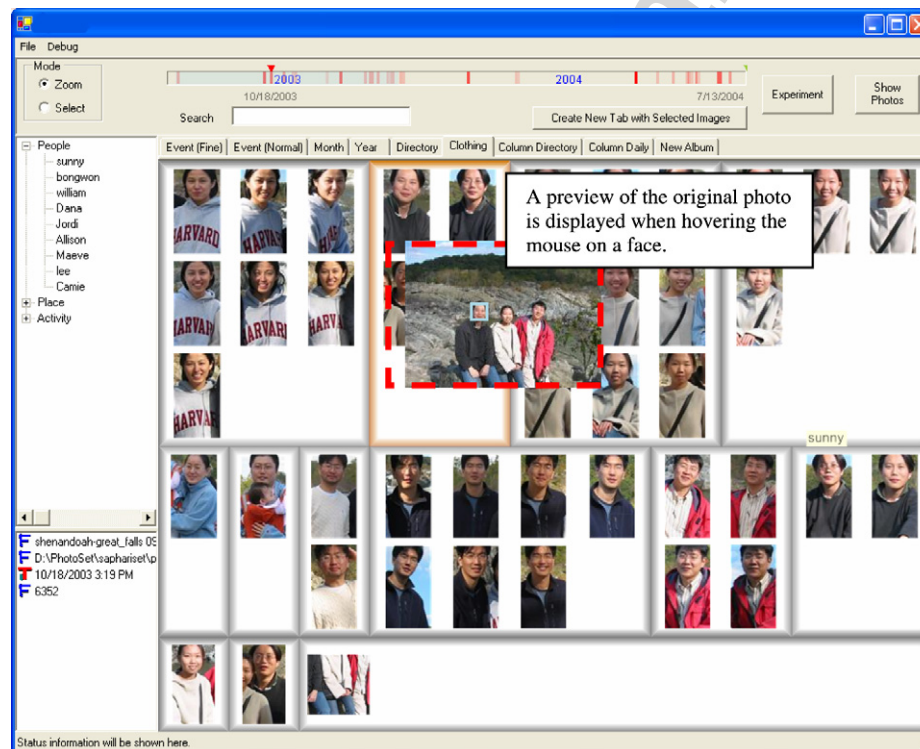


Fig. 8. People in photos are cropped, normalized, and laid out on the screen grouped by their clothing similarities. People who wear similar clothing are clustered together.

SAPHARI also offers various utility functions for manual regrouping. Users are provided with a context menu containing menu items such as “create a new face group”, “remove a face group”, “remove annotation”, “remove an unnecessary face image”, and “Not a Person” menu item. Due to the errors in face detection, sometimes, non-face images are recognized as a face. Clicking “Not a Person” removes the image from the face group. By using these functions, along with drag and drop techniques, users can relocate misclassified faces to where they belong.

6. User study

A set of user studies were conducted to investigate the effectiveness of semi-automatic annotation strategies applied to personal photo collections, and also to observe the annotation strategies users employed. The user study was divided into two parts. First, we observed and measured how event-based clustering effects participants’ annotation. Participants were asked to examine automatically identified event groups that are laid out in a 2D zoomable

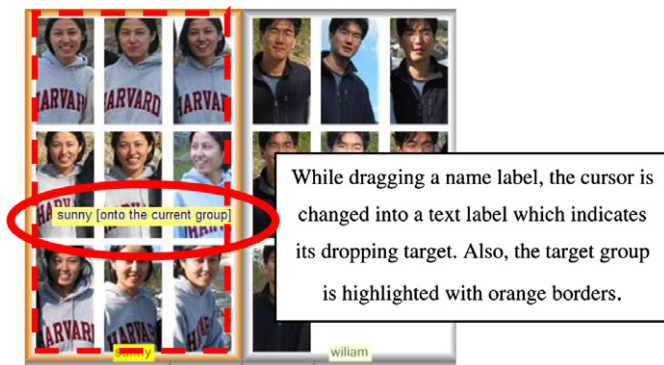


Fig. 9. Make bulk annotations by drag-and-dropping a name label on a face group.

space and to annotate them with some given key events (Event Task). Second, we compared annotation using clothing based person recognition with manual annotation. Participants were asked to identify people in a set of photos and to annotate them with appropriate name labels (Face Task).

While the task completion time was measured for comparison, the user study was not controlled. A controlled experiment requires the condition of each task to be identical to each other, which is not true in this user study. Rather than using a fixed controlled photo collection, we used participants' own photo collection to provide more realistic tasks by taking advantage of the participants' personal knowledge of their photos.

The study results showed some interesting patterns that provide valuable insight about semi-automatic annotation techniques. In addition, we were also able to observe various behaviors from participants while they were using SAPHARI.

6.1. Participants

There were seven participants in this study. All participants were computer science graduate students at the Uni-

versity of Maryland. There were four men and three women. All participants were familiar with computers. The summary of their photo collections are list as shown in Table 1.

Each participant was asked to provide more than five hundred photographs which had been taken over more than a six month period. It was difficult to recruit participants who were willing to provide their personal photos. Some of them were very concerned about their privacy especially because some of the photos had sensitive contents.

6.2. Method

A few days before meeting with participants, we asked them to fill out a pre-user study questionnaire. From those questions, we collected information about the events and people in the participants' photo collections. The list of events and people was used in the annotation tasks, which will be explained in later sections. On the day of the user study, we began the meeting by explaining the functions of SAPHARI. We gave details about navigation through a zoomable space and clarified the meaning of groups on the screen. A couple of sample browsing and annotation tasks were provided.

Table 1
Participant information

Participant	Digital photo experience	Estimated total size of the collection	Size of the collection provided for the study
P1	Three years	600	579
P2	Three years	2000	1245
P3	Four years	2000	1664
P4	Three years	1000	727
P5	2.5 Years	1000	464
P6	Three years	3000	1309
P7	Two years	1000	758

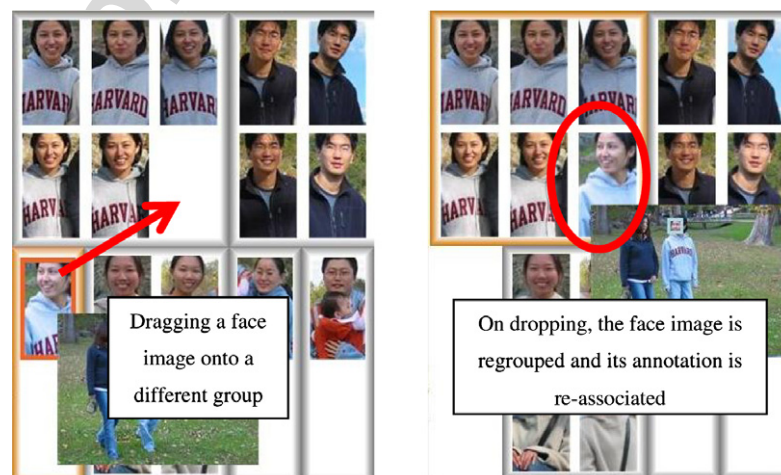


Fig. 10. Fix a misclassified face image. Moving a face image into a different face group updates the association between the face image and the name of the person.

Table 2 shows the task design. The user study employed a 2 × 2 design, with annotation technique and task type as independent variables. We measured the time per completion and the number of annotated items as dependent variables. We also collected some statistics about the photo collections such as the number of event groups, and the number of face groups. All participants were asked to finish the Event Task first, followed by the Face Task. However, the order of annotation techniques was counter-balanced to minimize learning effects. The total duration of the user study for each participant was about one hour. Participants were encouraged to “talk-aloud” while performing given tasks but not enforced. Upon finishing each task, we interviewed participants to gain more insight about their behaviors.

6.3. Event task

In the Event Task, participants were asked to annotate a set of specific events and we measured the task completion time. This task was intended to measure the efficiency and usability of the event based grouping compared to participants’ own folder-based photo organization. Through the pre-user study questionnaire, we found that all participants were using directories or folders to organize their photographs.

Participants were provided with two different types of interfaces: (1) a semi-automatic annotation interface where photos are grouped by automatically identified events and are laid out in a zoomable space (ESA); and (2) a manual annotation interface where photos are grouped by their

directory structure and are laid out in a non-zoomable canvas equipped with a vertical scrollbar (EM) as in Fig. 11. A recent user study (Cockburn et al., 2006) suggests that people can perform better with zoomable user interfaces when searching a document. Our research motivation was to explore novel user interfaces for personal photo management and we, thus, used state-of-the-art techniques including zoomable user interface and semi-automatic annotation methods. However, due to this design choice, factors other than pure semi-automatic annotation strategy influenced the participants’ annotation performance in the event tasks. Under the user study we designed, the effect of event based clustering cannot be separable from the benefit of the zoomable user interface.

As mentioned in the previous sections, participants had been asked to fill in the question, “Please state at least five interesting events or places in your photo collection”. Among the participants’ event list we collected, we randomly chose four events per participant for this task. Therefore, each participant was given different event tasks. For a controlled user study, a limited set of photos and limited target events should have been employed. However, with this non-controlled study design, the tasks are more consistent with real life situations.

Among the chosen four events, we assigned two events for the semi-automatic annotation interface (ESA) and the other two events for manual annotation interface (EM). In each task, participants were asked to annotate any number of photos that matched the given event. The order of tasks was counterbalanced. Half of the participants finished the semi-automatic annotation tasks (ESA)

Table 2
Four types of tasks were designed to compare the semi-automatic annotation strategy with conventional manual annotation approaches

Annotation strategy	Task type	
	Event annotation	Face annotation
Semi-automatic annotation	Photographs are laid on a 2D zoomable space grouped by events that have been automatically identified by the system. (ESA)	Cropped portraits of people are grouped by the similarity of clothing they wear (FSA)
Manual annotation	Photographs are laid out in a scrollbar canvas grouped by their directory structures (EM)	Photographs are laid out in a scrollbar canvas ordered by the date on which they were taken (FM)

Each task condition is denoted by the label; ESA, FSA, EM, and FM.

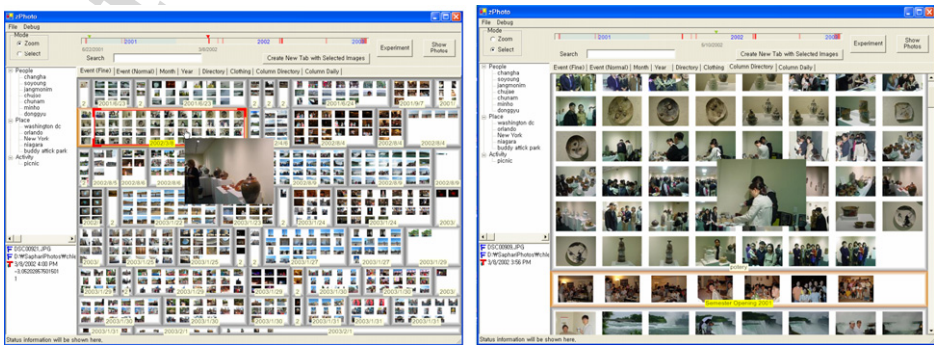


Fig. 11. Event tasks with two different settings. Left: photographs are grouped by events which have been automatically identified by SAPHARI. Right: photos are laid out by using participants own directory structure.

first followed by the manual annotation tasks (EM). The other half was asked to begin with the manual annotation tasks.

With the semi-automatic interface (ESA), participants were reminded that they can make annotations on either automatically identified photo groups or any ad-hoc photo groups. At any time, they were allowed to choose any number of photos as an annotation target. Such an annotation target can be selected by: (1) clicking with the control key pressed to add the clicked photograph to the current selection group; (2) clicking with the shift key pressed to add a range of photographs to the current selection group; or (3) selecting photographs by a marquee rectangle. Participants were allowed to drag a label onto a group of selected photos to annotate them at once. A single photo was also valid as an annotation target. The same ad-hoc photo grouping was also available in the manual annotation interface (EM).

Participants were reminded that there were two levels of event grouping, “Regular” and “Fine” grouping in ESA interface, and they were allowed to freely switch their views between them.

For each task, we recorded the completion time, the number of annotated photos, the number of label dragging and the number of event groups. The study administrator also took memos on the participant’s annotation and navigation strategies.

6.4. Face task

The second part of the user study was designed to measure the effectiveness of clothing based person recognition when applied to personal photo collections. Participants were asked to annotate a given set of photos with name labels as quickly and accurately as possible. Users were provided with two types of interfaces: (1) a semi-automatic annotation interface where cropped faces are grouped by clothing based person recognition (FSA); and (2) a manual annotation interface where non-cropped original photos are laid out on a canvas with scroll bars (FM) as shown in Fig. 12.

With the semi-automatic annotation (FSA), participants were allowed to make annotation only on face groups. Users were asked to drag a name label on a target face group, which annotates all faces in the face group. However, with the manual annotation interface (FM), participants were allowed to use the ad-hoc grouping with which users can select any number of photos as an annotation target.

Each participant was given four different face annotation tasks. The face annotation task was given with the instruction, “Please annotate the photos taken on [a specific sample date] with [a list of person]”. The task was easy to understand and realistic because participants were asked to make annotations on their own photos with people they were familiar with.

Before meeting with participants for the user study, two days from the photo collection per participant were manually picked for this task. Since the goal of the face task was to examine the feasibility and usefulness of clothing based person recognition, we needed to choose tasks that made sense. Among the dates from which participants took photos, we picked two dates per participant where: (1) at least three people appeared in the photos taken during a day; and (2) at least ten photos were taken in a day. However, once the dates are chosen, we used all photos taken on the dates to make annotation tasks more realistic.

As with the event information, a list of common people in the photos was extracted from the pre-study questionnaire with:

“Please state at least five people appear in your photo collection. You don’t have to list all the people. However please include people who are important to you – people who you want to find in your digital photo collection”.

The two picked dates and the two interface techniques generate four possible combinations for each participant: date one with clothing based annotation (FSA), date two with clothing based annotation (FSA), date one with manual annotation (FM), and date two with manual annotation (FM). Participants were asked to perform four tasks in a counterbalanced order. We recorded the completion

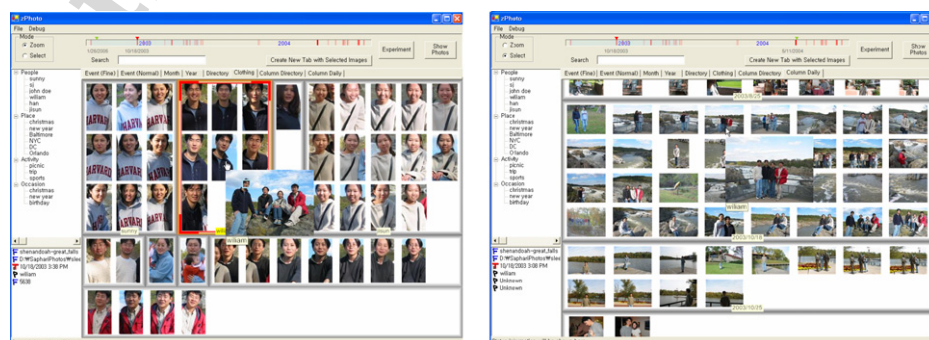


Fig. 12. Face annotation task interfaces. Participants were asked to annotate people in photos with two different interfaces. Left: clothing based annotation (FSA). Right: manual annotation (FM).

time, the number of annotated faces, the number of face groups, and the number of face regrouping. We also observed participants' annotation strategies.

6.5. Event task result

As shown in Fig. 13 and Table 3, the participants performed much better with the semi-automatic interface (ESA). When the automatic event grouping matched users' intentions, users were able to easily make bulk annotations on the photo groups. As shown in Fig. 13, the time per annotation was reduced 49% with the semi-automatic annotation (ESA), 0.367 s compared to the manual annotation (EM), 0.720 s.

A paired sample *t*-test was conducted on the time per annotation and there was a significant statistical difference depending on the annotation techniques, $t(1, 6) = 3.16$, $p = 0.019$.

During the user study, all participants complained about the repetition of selecting photos and drag-and-drops. One participant reported that "Sometimes, it's very difficult to select a group of pictures, especially if they are not adjacent". Especially with the manual annotation (EM), even

though participants were trying to make bulk-annotations as much as possible, selecting multiple annotation targets required significant effort from the users, such as clicking multiple times with shift key pressed, and using scrollbars. On the other hand, participants took advantage of event groups with semi-automatic annotation interface (ESA) where participants were allowed to make annotations on pre-clustered event groups. Participants spent less time to select annotation targets.

Overall, participants were positive about automatically identified events in the semi-automatic annotation interface (ESA). They immediately noticed the meaning of each event. One participant said, "This is Thanksgiving dinner and this is Christmas. And this is when my parents were here". He was very satisfied with the automatic event groups and reported, "This grouping is much better than my directories".

However, we also found that the efficiency of locating a specific photo was enhanced by zoomable user interfaces. Even though photos were represented as small thumbnails in ESA tasks, users were able to easily identify an individual photo through a small thumbnail, which eventually helps users to identify photo groups. However, the effect

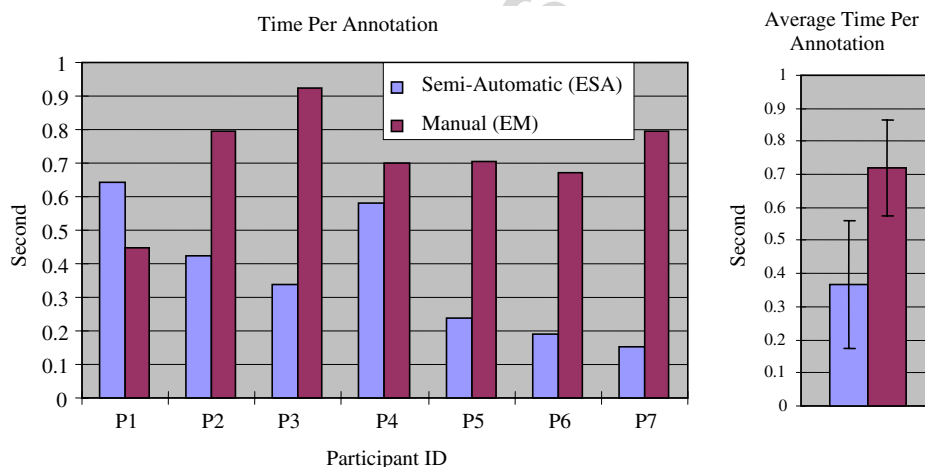


Fig. 13. Time per annotation from the event tasks. The left figure shows individual performance of participants and the right figure shows the average and the standard deviation of time spent per annotation.

Table 3
Annotation performance for the event tasks

Participants	Number of annotated photos and time per annotation in seconds			
	Semi-automatic interface (ESA)		Manual interface (EM)	
	Task 1	Task 2	Task 3	Task 4
P1	93 (0.473)	78 (0.843)	251 (0.408)	162 (0.511)
P2	100 (0.822)	215 (0.255)	97 (1.867)	193 (0.255)
P3	73 (0.369)	135 (0.323)	91 (0.907)	165 (0.930)
P4	89 (0.407)	89 (0.757)	97 (0.572)	34 (1.073)
P5	72 (0.752)	256 (0.094)	152 (0.279)	67 (1.673)
P6	143 (0.195)	58 (0.184)	19 (1.439)	72 (0.466)
P7	281 (0.127)	400 (0.169)	56 (0.642)	38 (1.022)

of zoomable user interface was not measured separately and the reported users' performance enhancement is combination of the two effects of event based clustering and zoomable user interface. We found it hard to determine if either one of them is the dominant factor in the performance improvement. Further study will be required to determine how each factor contributes to the difference.

SAPHARI provided two event granularities; "Regular" and "Fine" (see Section 4). Table 4 shows the grouping result. As the above table shows, "Regular" grouping provides a smaller number of large event groups while "Fine" grouping generates a larger number of tightly coupled photo groups. When asked which event grouping was better, six out of seven participants answered that they preferred "Regular" grouping. Participants did not care much about detailed events. Since they could remember most events in their photo collections, they preferred to find a high level event and then do a visual search among the photos in that event. During the event task, only one participant (P2) used the "Fine" event grouping. In addition, we found that participants used "Regular" grouping during the Event Task because the events used in the user study were also more suited for "Regular" grouping. For example, a coarse event such as "Parents Visit" was well-covered by "Regular" grouping while more detailed events such as "Family Dinner (during the parent visit)" could be detected by "Fine" grouping. Most of the participants' events, collected in the pre-study questionnaire, were bigger units such as "Parent Visit" rather than detailed events, and therefore most participants were satisfied with the "Regular" event grouping during the user study.

During the user study, some problems were also observed with the semi-automatic annotation interface (ESA). One participant in particular had problems with event groups. Some photos in her collection were altered when rotated and shrunk. The EXIF headers of the photos were destroyed and the photos' timestamp showed the date of the modification, not the actual date the photo was taken. Due to this problem, some events were grouped incorrectly and she had to unscramble the spoiled event groups. In addition, four participants also reported there were errors in event boundaries (ESA). However, rather than fixing the boundaries, participants used ad-hoc grouping to select annotation targets. Because the photos are laid out in chronological order, they easily identified appropriate

annotation targets and selected them for making annotations. In our user study, only two participants (P1 and P2) fixed the event boundaries (one time and two times, respectively) even though there are more event boundary detection errors.

6.6. Face task result

The goal of the face task was to investigate the potential benefit of clothing based person recognition. With the Face Task, there was only a 6% improvement with the semi-automatic annotation with clothing based person recognition (FSA, 2.54 s per annotation) compared to the manual annotation (FM, 2.71 s per annotation). As shown in Fig. 14 and Table 5, the time per annotation did not show significant difference.

Overall, participants spent a significantly longer time to annotate faces than events. While performing manual annotation tasks, one participant reacted that "It requires too much effort to do face annotation even though it does not seem to take much time". Another participant even complained about wrist fatigue after finishing the Face Task. After the user study, all participants agreed that the face annotation task was too time consuming especially because they had to select target photos one by one. In the Event Task, participants were able to take advantage of bulk annotation even with the manual annotation interface (EM) because the photos were grouped by the directory. However, with manual annotation on the Face Task (FM), participants were not able to make bulk annotation easily. Photographs that contain a specific person are not necessarily located together on the screen. They were scattered on the screen and participants were required to identify the people in photos one by one. Even though there was no significant difference in the task completion time, participants clearly became more tired with the manual annotation strategy.

We observed a few interesting patterns when participants were making annotations using the clothing based annotation (FSA). First, due to inaccurate results from the face detection algorithm, there were a quite number of non-faces that were recognized as faces. It caused SAPHARI to include those non-face images in the face groups. Participants spent some effort to remove those non-face images from face groups. This procedure slowed

Table 4
Event based clustering result

Participant	Collection size	Number of event group with regular grouping	Number of event group with fine grouping	Directories in photo collection
P1	579	22	60	20
P2	1245	20	37	1
P3	1664	52	109	38
P4	727	130	243	8
P5	464	84	141	67
P6	1309	34	91	17
P7	758	25	99	16

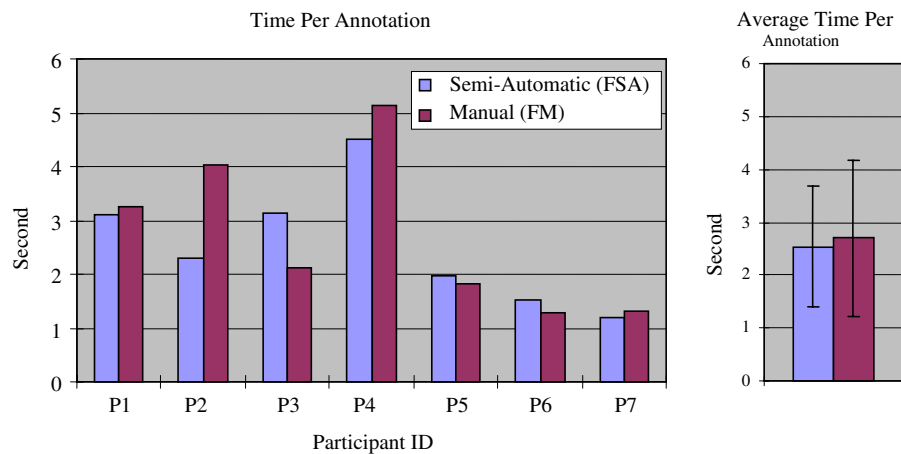


Fig. 14. Time per annotation for the face task. The left chart shows individual performance of participants and the right shows the average and the standard deviation of time spent per annotation.

Table 5
Annotation Performance for the Face Tasks

Participants	Number of annotated faces and time per annotation in seconds			
	Semi-automatic (FSA)		Manual (FM)	
	Task 1	Task 2	Task 1	Task 2
P1	19 (4.48)	19 (1.70)	42 (2.92)	82 (3.40)
P2	18 (2.06)	23 (2.47)	21 (3.72)	34 (4.22)
P3	28 (3.04)	12 (3.39)	65 (2.36)	24 (1.50)
P4	11 (5.11)	21 (4.19)	36 (5.93)	50 (4.55)
P5	33 (2.17)	29 (1.76)	64 (1.92)	69 (1.73)
P6	35 (1.27)	19 (1.98)	49 (1.26)	38 (1.32)
P7	36 (0.65)	65 (1.51)	37 (1.06)	65 (1.44)

down the annotation process. Second, participants spent some time organizing non-annotation-target people while performing the Face Task. While the Face Task asked participants to make annotations only with the given people, in many cases, participants also tried to organize people who were not annotation targets. For example, one participant was asked to annotate only her family members. However, when the given Face Task also showed a lot of images of her friends, she also identified and organized some of her friends as well as her family members. The study results only count annotated family members excluding all other identified persons.

In addition, we observed that participants spent some time just looking at face images with the semi-automatic annotation interface (FSA) before performing any annotation task. With the manual annotation interface (FM), participants immediately began to annotate. When asked what they were doing, participants responded that they were examining the results and planning how to group them. Participants also showed much more enthusiasm with the clothing based person recognition. One participant was annotating photos which were not in the given task. He stated that he just wanted to annotate everybody in the collection.

As shown in Table 6, participants corrected a significant number of recognition errors while performing the Face

Tasks. To annotate 368 faces (the sum of all annotated faces; the last column in Table 6), they made 155 corrections, which indicates that the accuracy of the clothing based recognition is 58%. In Face Recognition Vendor Test (FVRT) in 2002 (Phillips et al., 2003), the average recognition accuracy of the best three systems for outdoor face images were 51% (the best system with 54% accuracy). Even though the experiment conditions are different from each other, the result of clothing based person recognition is very encouraging compared to systems evaluated in FRVT 2002. SAPHARI is able to recognize humans without using any facial features. We believe that our clothing based recognition technique has the potential to increase the accuracy of face based human recognition systems.

6.7. Subjective satisfaction

Through the post-user study questionnaire, much stronger differences emerged. After finishing the Face Task, participants answered questions about their satisfaction with the interfaces they used in the study. Fig. 15 shows the average and standard deviation scores on a seven point scale (1 = disagree, 7 = agree) for participants' responses to a number of ease of use and preference ratings. A one-way analysis of variance (ANOVA) was run on each measure to test for differences between interface techniques.

Table 6
Clothing based person recognition result

Data set		Number of face group	Number of detected face	User's error correction	Number of face group after correction	Number of annotated face
P1	Set 1	6	19	6	6	19
	Set 2	25	40	10	14	19
P2	Set 1	45	146	5	46	18
	Set 2	18	47	11	15	23
P3	Set 1	20	54	9	10	28
	Set 2	10	26	11	3	12
P4	Set 1	15	30	16	9	11
	Set 2	17	39	16	9	21
P5	Set 1	14	46	16	4	33
	Set 2	18	44	10	10	29
P6	Set 1	16	36	11	4	35
	Set 2	17	41	13	11	19
P7	Set 1	7	36	3	7	36
	Set 2	23	66	18	10	65

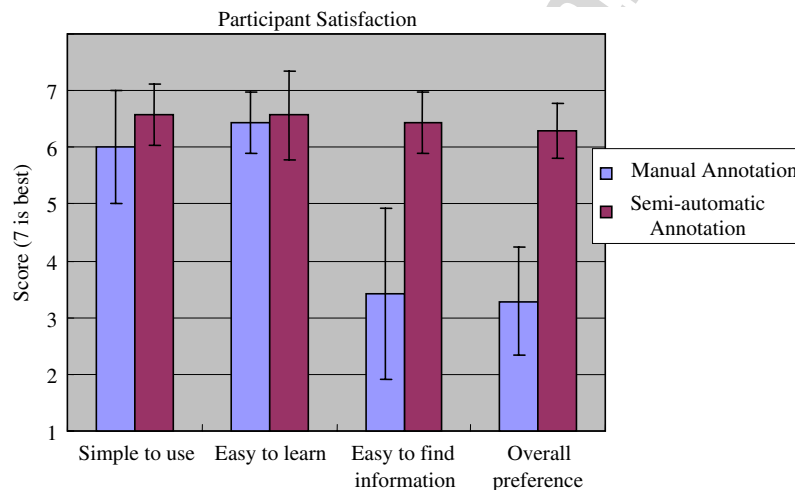


Fig. 15. Participants' subjective satisfaction was measured by the post-user study questionnaire.

There was no difference in the first two questions; “simple to use” and “easy to learn”. Participants reported that all annotation interfaces were easy to use. Participants immediately grasped the main concepts of annotation and event groups. They also had no trouble navigating using the zoomable interface. All participants were able to finish the given task successfully without any problem with two different interface techniques and they answered very positively for both interfaces when being asked about the ease of use and learnability.

However, the answer to the question, “easy to find information” showed a clear preference for the semi-automatic approaches. Participants were significantly positive with semi-automatic annotation interfaces compared to manual annotation interfaces, $t(12) = 6.74$, $p < 0.001$. Since the semi-automatic interfaces provided photos in a 2D zoomable space, participants were able to easily zoom in any photos. In addition, the zoomable interface also pro-

vided quick previewing of photos. With mouse hovering, the interface provided a preview (about 200×150 in pixels) of photos under the cursor. Most participants were able to take advantage of the zoomable user interface to navigate their photo collections. One participant mentioned that the previewing was “definitely useful”. This observation implies that zoomable user interface techniques clearly helped users to identify annotation targets. However, automatic event identification also helped users when making annotations. Participants took advantage of automatically identified event groups by making bulk annotations onto the identified photo group.

For the “overall satisfaction” question, participants unanimously preferred semi-automatic annotation interfaces with very strong statistical significance, $t(12) = 7.42$, $p < 0.001$. Response was very positive. A few participants were even interested in continuing to use SAPHARI on their personal computer.

7. Discussion

SAPHARI is designed based on a number of design principles we identified in our earlier studies (Bederson, 2001; Suh et al., 2003). First, SAPHARI facilitates bulk annotation by clustering photos into meaningful groups. In SAPHARI, photos can be grouped based on their timestamp so that each group corresponds to a meaningful event (see Section 4). Also, SAPHARI can group similar-looking people together so that users can label them easily (see Section 5). Second, SAPHARI is designed to provide clear feedback in response to users' input. In "Event" layout, users' correction on event boundaries immediately regroups photos on the screen. In "Clothing" layout, moving a face image between face groups gives very clear feedback (Fig. 10). Third, in SAPHARI, users can always override suggestions from automatic recognition algorithms. Event boundaries and person label association can be easily overridden. Fourth, SAPHARI takes advantage of timestamps of photos to provide contextual information. In "Event" layout, photos are laid out based on their chronological orders in a zoomable space (i.e., Fig. 5), and, thus, allowing users to naturally browse photos around a specific photo. We found this feature especially useful when a user wants to update event boundaries. Lastly, users are not forced to make annotations at any time. Annotations can be made incrementally and interactively.

In SAPHARI, the semi-automatic annotation interface is integrated with a zoomable user interface. Participants were provided with zoomable navigation techniques, zooming in to see more detail and zooming out to see the overview of images. During the user study, we observed that the zoomable navigation helped participants when searching annotation targets. The participants were able to find events in their collection immediately in a zoomable space. One interesting characteristic of personal photo collections is that users are already very well aware of photos in their collection. Combined with zoomable user interface techniques, familiar photos seem to play a very important role in efficient browsing. Compared with our previous user studies in (Suh et al., 2003), which were designed to browse non-familiar images, search performance appeared to be improved when participants were browsing their personal photos. In the previous study, users spent about 30 seconds on average to find a single photo that contains a specific search target object among 100 photos. However, in this study, participants spent about 46 s to finish the event annotation task (ESA) with semi-automatic and zoomable user interface techniques. In the ESA task, there were 964 photos on the screen on average and participants dragged labels 3.9 times on average. Even though participants had to browse and locate much larger number of photos, they finished the given tasks reasonably fast. Although it is hard to directly compare the two studies, we believe that participants of this user study performed much better

when searching photos. We suspect it is because the participants were already familiar with their personal photos. Even when photos were represented in very small thumbnails, participants were able to remember details of the photos. This suggests that zoomable user interfaces, where small thumbnails are frequently used, have a great potential when used for handling familiar information. Even though zoomable user interfaces were not the focus of this study, this was an interesting and unexpected observation, so we report it here.

There are a number of possible improvements for the research presented in this paper. The face detector used in SAPHARI can be replaced with one with higher accuracy. The Viola–Jones face detector (Lienhart and Maydt, 2002; OpenCV, 2005; Viola and Jones, 2004) used in SAPHARI is often heavily affected by lighting conditions and fails to work properly. In addition, SAPHARI only detects frontal face views. Although the frontal view is the most common form of faces in photos, supporting lateral views will increase the accuracy of human recognition. More efficient face detector will increase the effectiveness of clothing based person recognition. In addition, overall human recognition accuracy can be improved if the clothing based approach is used in conjunction with other facial feature based techniques (Phillips et al., 2003). In addition to facial features, we expect using social structures suggested in (Naaman et al., 2005) can help to achieve more accurate human recognition.

As explained earlier, there are a couple of assumptions we made when designing SAPHARI; we assumed that: (1) photo collections are episodic; and (2) people usually wear the same (and unique) clothing within a day. When these assumptions are not met, SAPHARI does not work well. For example, when people wear uniforms or when people are in bathing suits, clothing based person recognition cannot be applied. SAPHARI also assumed all photos have valid timestamps. When the timestamps of photographs are modified, SAPHARI generates inaccurate results. Further research has to be performed on cases where these assumptions are not met.

There also were some usability issues with SAPHARI. Some participants did not like the drag-and-drop interaction. Some of them complained about difficulties in marquee-selecting images. Sometimes, users were confused between "selection mode" which allows users to select images and "zoom mode" which enable users to navigate a zoomable space. Further refinement is needed on these issues.

The user study design presented in this paper also has some limitations. During the user study, we measured task completion time to access users' annotation performance. However, we also allowed users to talk about what they were doing, which could have an effect on task completion time. In addition, the effect of zoomable user interface could have been captured more precisely if the user study had added a "zoomable user interface only" condition (without semi-automatic strategies).

8. Conclusion

In this paper, we explored semi-automatic techniques to help users make accurate annotations with low effort. While metadata is very important for browsing and searching photos, it is hard to acquire accurate metadata associated with photos. Automatic metadata extraction is typically fast but inaccurate, while manual annotation is slow but accurate. We designed and implemented a semi-automatic annotation prototype, SAPHARI which combines these two techniques. The prototype generates photo clusters which facilitate efficient bulk annotation. SAPHARI automatically creates photo clusters with hierarchical event clustering and clothing based person recognition.

We performed a series of user studies with seven participants. The results showed the potential benefit of the semi-automatic annotation when applied to personal photo collections. Users were able to make annotations 49% and 6% faster with the semi-automatic annotation interface on event and face tasks, respectively. Hierarchical event grouping strategy was able to provide users multiple sets of event groupings. Although six of out seven participants preferred the “Regular” event grouping with “Fine” grouping, we believe hierarchical event organization can be useful when users want to have alternative event groups. Clothing based person recognition achieved reasonable recognition accuracy. This strategy has a potential to increase human recognition performance when combined with other human recognition techniques.

Despite the lack of significant performance improvement in the Face Task, the study participants significantly preferred the semi-automatic approach. We believe that this is because the kind of work the participants did with the semi-automatic interface better matched their skills and interests. The manual interface required them to perform numerous repetitive and boring tasks which became mind-numbing over time. The semi-automatic interface, on the other hand, required the participants to work at a higher level – paying attention to the results of the automatic clustering algorithm, spotting and correcting mistakes. The tasks for this interface were fewer, but more complex.

So, not only do semi-automatic approaches offer the potential for better performance as the content analysis algorithms improve (which they are bound to do), but we also believe that semi-automatic approaches work better for people because they map more naturally to what people tend to be good at, thus providing a preferable experience.

Semi-automatic annotation is still in its early stage. Even though computer vision research has developed many useful techniques, only a few are directly applicable to personal photos. We hope further research will provide useful automatic recognition techniques which can be integrated with user interface strategies to help users manage ever-growing personal photo collections.

References

- Adobe Photoshop Album, 2006. Adobe Systems Incorporated., www.adobe.com/products/photoshopalbum.
- Barron, C., Kakadiaris, I., 2001. Estimating anthropometry and pose from a single image. *Comput. Vis. Image Und.* 81 (3), 269–284.
- Bederson, B.B. 2001. PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps. *UIST 2001, ACM Symposium on User Interface Software and Technology*, CHI Letters, ACM Press, 3(2), 71–80.
- Bhattacharya, A., 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Maths Soc.* 35, 99–110.
- Cockburn, C., Gutwin, C., Alexander, J. 2006. Faster document navigation with space filling thumbnails. In: *Proceedings of the ACM CHI 2006 Conference on Human Factors in Computer Systems*. ACM Press, pp. 1–10.
- Cooper, M., Foote, J., Girgensohn, A., Wilcox, L. 2003. Temporal Event Clustering for Digital Photo Collections. In: *Proceedings of the 11th ACM International Conference on Multimedia*, (MM 03). ACM, 1 (3), pp. 269–288.
- Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W., 1994. Efficient and effective querying by image content. *J. Intell. Inf. Syst.* 3 (3–4), 231–262.
- Flickr, 2005. Yahoo! Incorporated, www.flickr.com.
- Gargi, Ullas. 2003. Consumer Media Capture: Time-Based Analysis and Clustering, Hewlett-Packard Tech. Report HPL-2003-165.
- Girgensohn, A., Adcock, J., Cooper, M., Foote, J., Wilcox, L. 2003. Simplifying the Management of Large Photo Collections. *Human-Computer Interaction*, INTERACT 03, IOS Press, 196–203.
- Girgensohn, A., Adcock, J., Wilcox, L. 2004. Leveraging face recognition technology to find and organize photos. In: *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, ACM Press, pp. 99–106.
- Graham, A., Garcia-Molina, H., Paepcke, A., Winograd, T., 2002. Time as essence for photo browsing through personal digital libraries. In: *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM Press, pp. 326–335.
- JEITA (Japan electronic and information technology industries association). (2002). Exchangeable image file format for digital still cameras: Exif Version 2.2 www.exif.org.
- Kang, H., Bederson, B.B., & Suh, B. (accepted for publication). Innovative GUIs for Dynamic Digital Image Management. *Int. J. Human-Computer Interact.*, Lawrence Erlbaum Associates.
- Kang, H., 2003. Personal media exploration with semantic regions. In: *Extended Abstracts on Human factors in computing systems (CHI 2003)*. ACM Press, pp. 668–669.
- Kang, H., Shneiderman, B., 2000. Visualization methods for personal photo collections: browsing and searching in the photoFinder. *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2000)*, vol. 3. IEEE, pp. 1539–1542.
- Koenemann, J., Belkin, N.J., 1996. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In: *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*. ACM Press, pp. 205–212.
- Kohonen, T., 1989. *Self-Organization and Associative Memory*. Springer-Verlag.
- Kuchinsky, A., Pering, C., Creech, M.L., Freeze, D., Serra, B., Gwizdka, J., 1999. FotoFile: a consumer multimedia organization and retrieval system. In: *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI 99)*. ACM Press, pp. 496–503.
- Leung, L., Zhang, J.S., Xu, Z.B. 2000. Clustering by Scale-space Filtering, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 22(12), IEEE, 1396–1410.
- Lienhart, R., Maydt, J. 2002. An Extended Set of Haar-like Features for Rapid Object Detection. *IEEE ICIP 2002*, IEEE, vol. 1, 900–903.

- Loui, A., Savakis, A., 2000. Automatic image event segmentation and quality screening for albuming applications. In: *Proceedings of IEEE International Conference on Multimedia and Expo*. IEEE, pp. 1125–1128.
- Naaman, M., Yeh, R.B., Garcia-Molina, H., Paepcke, A., 2005. Leveraging context to resolve identity in photo albums. In: *Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2005)*. ACM, pp. 178–187.
- Naaman, M., Song, Y.J., Paepcke, A., Garcia-Molina, H., 2004. Automatic organization for digital photographs with geographic coordinates. In: *Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2004)*. ACM, pp. 53–62.
- Open CV, Open Source Computer Vision Library, Intel Inc. www.intel.com/research/mrl/research/opencv.
- Phillips, P.J., Grother, P.J., Michaels, R.J., Blackburn, D.M., Tabassi, E., Bone, J.M. 2003. Face recognition vendor test (2002): Evaluation report, NISTIR 6965, NIST, Gaithersburg, MD.
- Picasa, 2006. Google Inc. www.picasa.com.
- Platt, J.C., Czerwinski, M., Field, B. 2002. PhotoTOC: Automatic Clustering for Browsing Personal Photographs, Microsoft Research Technical Report MSR-TR-2002-17.
- Rodden, K., Wood, K., 2003. How do people manage their digital photographs? In: *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI 2003)*. ACM Press, pp. 409–416.
- Rosales, R., Sclaroff, S., 2000. Inferring body pose without tracking body parts. *Proceedings of IEEE Computer Vision and Pattern Recognition*, vol. 2. IEEE, pp. 721–727.
- Riya.com, 2005. Ojos Incorporated. www.riya.com.
- Shen, C., Lesh, N., Moghaddam, B., Beardsley, P., Bardsley, R.S., 2001. Personal digital historian: user interface design. In: *Extended Abstract of CHI 2001*. ACM Press, pp. 29–30.
- Shen, H.T., Chin, O.B., Tan, K., 2000. Giving meanings to WWW images. In: *Proceedings of Eight ACM Multimedia Conference (SIGMM 2000)*. ACM Press, pp. 39–47.
- Silverman, B.W., 1986. Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*, vol. 26. Chapman & Hall.
- Tuffield, M., Harris, S., Dupplaw, D.P., Chakravarthy, A., Brewster, C., Gibbins, N., O'Hara, K., Ciravegna, F., Sleeman, D., Wilks, Y., Shadbolt, N.R. 2006. Image Annotation with Photocopain, <http://eprints.ecs.soton.ac.uk/12004>.
- Suh, B., Ling, H., Bederson, B.B., Jacobs, D.W. 2003. Automatic Thumbnail Cropping and Its Effectiveness, *UIST 2003, CHI Letters*, 5(2), ACM Press, 94–104.
- Viola, P., Jones, M., 2004. Robust real-time object detection. *Int. J. Comput. Vis.* 57 (2), 137–154.
- Von Ahn, L., Liu, R., Blum, M., 2006. Peekaboom: a game for locating objects in images. In: *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI 2006)*. ACM Press, pp. 55–64.
- Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., Field, B., 2001. Semi-automatic image annotation. In: Hirose, M. (Ed.), *Proceedings of Human-Computer Interaction – Interact '01*. IOS Press, pp. 326–333.
- Yang, M., Kriegman, D., Ahuja, N., 2002. Detecting faces in images: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (1), 34–58.
- Yoshitaka, A., Ichikawa, T., 1999. A survey on content-based retrieval for multimedia databases. *IEEE Trans. Knowl. Data. Eng.* 11 (1), 81–93.
- Zhao, W., Chellappa, R., Philips, P.J., Rosenfeld, A., 2003. Face recognition: a literature survey. *ACM comput. surv.* 35 (4), 399–458.