Semi-Automatic Image Annotation Using Event and Torso Identification

Bongwon Suh, Benjamin B. Bederson
Department of Computer Science,
Human-Computer Interaction Laboratory
University of Maryland
College Park, MD 20742 USA
+1 301-405-2764
{sbw, bederson}@cs.umd.edu

ABSTRACT

Annotation is important for personal photo collections because acquired metadata plays a crucial role in image management and retrieval. Bulk annotation, where multiple images are annotated at once, is a desired feature for image management tools because it reduces users' burden when making annotations. This paper describes an approach for automatically creating meaningful image clusters for efficient bulk annotation. These techniques are not perfect and so are integrated into a bulk annotation interface where users can manually correct errors. We present hierarchical event clustering and torso based human identification techniques. Hierarchical event clustering provides multiple levels of "event" groups. For identifying people in images, we introduce a new technique which uses torso information rather than human facial features.

Keywords

Image browsing, annotation, clustering, event identification, computer vision, face detection, zoomable user interface

INTRODUCTION

As the volume of images one person needs to handle increases, it is more difficult to manage them on computers. There is, therefore, a demand for image management systems for efficient image organization, search, and browsing. However, personal photos taken by consumer digital cameras typically include only basic metadata such as time and date, which are often not enough. For example, event information that users may care about, such as "Birthday Party", has to be acquired from users.

There are roughly three ways of acquiring metadata for images. They are 1) automatic extraction by image analysis; 2) manual annotation; and 3) semi-automatic annotation such as suggested in [15]. Automatic metadata extraction by image analysis is typically fast – compared to manual annotation - but inaccurate while manual annotation is slow but accurate. The semi-automatic annotation is a combined technique of the two approaches. By using relevance feedback from users, initial metadata which has been obtained automatically are updated incrementally. With appropriate interfaces, semi-automatic annotation has the potential of having the highest overall performance.

Annotation is defined as a process which involves labeling the semantic content of images (or objects in images) with a set of keywords or semantic information. Annotated information is very important for image retrieval since it allows keyword-based search and helps organizing photos. However, manual annotation is usually labor intensive and tedious. 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, although individual placement might still be needed [7][8]. Therefore, it is important to provide meaningful image groups to users, ready to be annotated, to facilitate bulk annotation which can reduce users' burden. A recent user study [13] analyzed casual users' usage patterns on personal image collections. They found that "event" and "person" are the most frequently used metadata as well as chronological order of photos. Thus, it is mandatory for image management tools to support those types of metadata.

In this paper, we introduce two strategies to identify meaningful annotation units. Our first method is to cluster photos by using timestamps so that each group of images can correspond to a meaningful "event" as shown in Figure 1. Our second approach is to use a torso based human recognition technique to group similar persons together. Users can drag labels onto those automatically identified images groups to complete annotation.

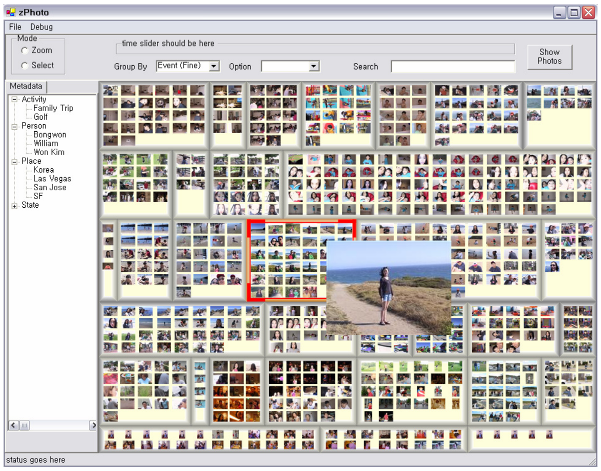


Figure 1 A prototype interface for semi-automatic image annotation based on PhotoMesa[2]. Photos are laid out within a zoomable space grouped by event that is detected by hierarchical event clustering. Users can drag labels on image groups to make bulk annotation.

RELATED WORK

There has been much research to simplify the image annotation process. QBIC [5] tried to use image-based analysis techniques to extract metadata, but automatic feature extraction is still not very accurate or robust. People are the only source that can add reliable, accurate and appropriate information to images. However, it is a burden for users to annotate metadata on images. Kang et al. [7] developed a direct annotation method that focuses on labeling names of people in photos. Similarly, Adobe PhotoShop Album [1] 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. While it saves users typing work, users still have to perform drag and drop many times.

Wenyin et al. [15] introduced an approach to semi-automatically and progressively annotating images with keywords. The progressive annotation process is embedded in the course of integrated keyword-based and content-based image retrieval and user feedback. The strategy of semi-automatic image annotation was found to be better than manual annotation in terms of efficiency and better than automatic annotation in terms of accuracy. But as stated in [15], the prototype needs a more refined user interface strategy.

Also, there are many approaches to automatic event clustering for digital photo collections. Cooper et al. introduced a temporal similarity-based approach [3]. Platt et al. [12] uses an adaptive local threshold method to detect event boundaries. While some of them work pretty well in terms of precision and recall measure, most of them did not focus on internal event hierarchies, which will be explained later.

There is a vast number of computer vision techniques to detect and recognize human faces [17], and some representative approaches are described here. Nakajima et al. address surveillance scenarios where the pose of people is unconstrained which makes it difficult to apply common face recognition algorithms [10]. They capture human body as a model and use SVM to differentiate human bodies based on their color features. Srihari et al. [14] uses face detection and low-level features of the background as a tool for a content-based retrieval system and they did not use it for identifying humans. Kuchinsky et al. [8] also used facial information for image management.

HIERARCHICAL EVENT CLUSTERING

As stated earlier, "event" is one of the most important units for personal image organization. Much research has been performed to find meaningful event clusters from image collections [2][6][12]. One of the interesting characteristics of personal image collections is that personal images are typically sporadic or episodic in terms of temporal order [6]. Users don't usually take photos on a regular basis, such as one shot a day. When there is a special occasion or when a user is carrying a camera, they take many shots in a relatively short period of time. Then, there is a pause until the next series of shots. Users usually remember the occasions or events and like to search or browse images by them [13].

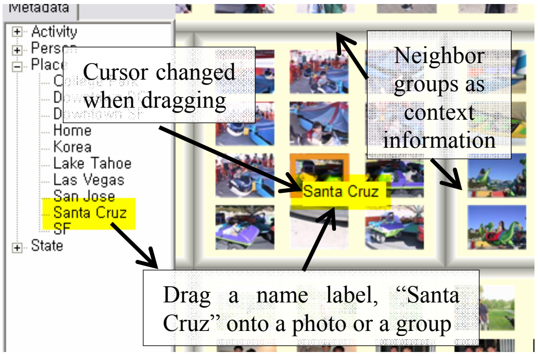
In addition, we found another interesting pattern in event clusters. The images in a personal collection usually have a temporal hierarchy. For example, "Summer Camping Trip" which spans on June 13th - 17th, can contain many subordinate event units such as "Hiking" on 14th, "Canoeing" on 15th, and so on. We found that users want to identify each separate event as well as "Camping Trip" as a whole. These event units are important because they become units for making annotation.

We introduce multiple levels of event clusters for image annotation. Our prototype allows users to choose different levels for annotation. When users want to annotate a broad event such as "Summer Camping Trip", the interface provides coarsely clustered groups. And, when users want finer granularity events such as "Hiking" on 14th, more tightly clustered groups are provided. Hierarchical event clustering provides a more flexible way to annotate images compared to fixed event clustering techniques.

However, automatic recognition inevitably introduces incorrectly clustered images [3]. When users find the recognized cluster inappropriate even with the right granularity, it is usually because some images in a cluster should have been included in one of its neighboring groups. In most cases, the errors – misclustering – can be corrected by resetting event boundaries. Our prototype provides an intuitive and natural way to fix the errors. Users can quickly merge two neighboring groups or move an individual photo to other groups by drag-and-dropping as well. In addition, users have the freedom to split one event group into multiple groups. One of the features that we see as being particularly important here is that the surrounding group information should also be provided to users. Our prototype enables users to see neighboring groups naturally as context information by showing groups in a zoomable space. The zoomable user interface technique of the prototype lets users to see multiple groups at once without losing focus on the current group [2].

While the prototype provides multiple event levels, the logical structure between events should also be maintained. As users change the original clustering that had been automatically identified, we found that two conditions should be kept to maintain the logical integrity of the event hierarchy. They are: 1) when events are merged at a finer level, the event groups cannot be split at coarser levels; 2) When events are split at a coarser level, those events cannot be merged in finer levels. By using rules, users' intentions propagate to all levels so that more efficient annotation is possible. For example, when users split a single cluster into two groups at a coarse level, the change is automatically applied to every finer level.

In our prototype, we used the algorithm in [12] for initial cluster recognition by giving different constant values to control the tightness of clustering. However, other algorithms such as Cooper et al. [3], also can be used to form hierarchical clusters.



Cursor changed when dragging

Neighbor groups as context information

Drag a name label, “Santa Cruz” onto a photo or a group

Figure 2 Annotation by drag-and-drop. Users can drag a name label onto an image or a group to annotate photos. With the shift key pressed, the prototype chooses a group as its drop target instead of a single image under the mouse cursor. When a label is dropped on a group, photos in that group are annotated with the label at once (Bulk annotation).

Users can use the prototype for just browsing photos without making annotations at all. They can browse images grouped by events which were identified automatically as shown in Figure 1. Besides hierarchical event clustering, the prototype also provides various photo groupings. Photos can be laid out on the screen grouped by month, year, or directories which hold image files. When users choose to add captions or fix event boundaries, they can begin the annotation process as shown in Figure 2.

HUMAN IDENTIFICATION BY TORSO ANALYSIS

Along with the event information, the people in photographs are crucial for finding photos of interest [7] [8]. Face is the most crucial information for identifying people. However, research about recognizing human faces have had limited success and face recognition in an uncontrolled environment is still very challenging. For example, even for the best face recognition systems, the recognition rate for faces captured outdoors, at a false accept rate of 1%, was only about 50% [11]. Thus, we concluded that we cannot rely solely on human face recognition to identify people in photos.

However, there are interesting patterns in personal photo collections, which can be used to tackle this problem. We have noticed that people have a tendency not to change their clothing within a day. And, as stated earlier, people tend to take many photos in a day when they carry a camera. For example, the data set used in [6] shows that photos are taken approximately one day out of ten. And, on the day they take photos, they take about twelve photos.

Our approach is based on the assumption that people who appear in photos taken in one day and wear similar clothing and are, in fact, the same person..

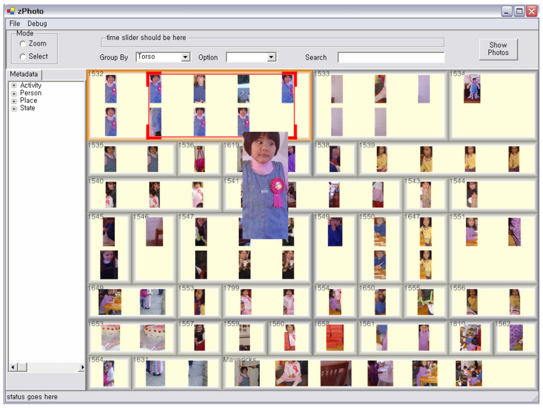


Figure 3 Identified people who are cropped from photos are laid out on the screen ready to be annotated. Torsos are detected by using face detection. Similar torsos which have short visual distances with each other are clustered together.

To model the clothing that people wear, we propose a technique which uses the visual features of torsos. The prototype identifies the torso as a region under a face. The system finds faces in photos by using the algorithm in [9] and identifies the torso as shown in Figure 4. The prototype clips out faces along with corresponding torsos from photos and lays them out on the screen as shown in Figure 3. With the prototype, users can make bulk annotations on people by dragging name labels on person groups. Also, recognition errors can be corrected by dragging human face images into the right group. This is motivated by the same principle that seeing multiple groups of photos at a time will enable more efficient correction as in the event clustering.

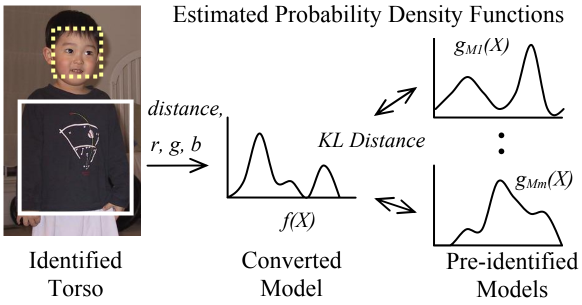


Figure 4 Human Torso Model.

The basic strategy of our torso analysis is to measure the visual distance between torsos. When a distance is below an empirically determined threshold, the system classifies the torsos as the same one. Each torso is represented with a probability density function by using Gaussian kernel density estimation [4]. A four dimensional feature vector $X = (distance, r, g, b)$ is employed to model the torso where *distance* is defined as the relative vertical position in the region and *r, g,* and *b* present color information. Compared to the cases that use only color features, adding spatial information, *distance* increases the recognition accuracy [16]. Suppose that $g_{mi}(X)$ is the density function of a predefined torso model and $f(X)$ is the density function of a test instance as shown in Figure 4. Then, the nearest model from $f(X)$ is determined by using *Kullback-Leibler Distance* [4] as:

$$\textit{Nearest Model from f(X)} = \underset{i}{\operatorname{argmin}}(KLDist(f(X), g_{Mi}(X)))$$

$$\text{,where } KLDist(f(X), g_{Mi}(X)) = \sum_{x \in X} f(x) \log \frac{f(x)}{g_{Mi}(x)} \quad [4]$$

Torso based analysis is robust, pose invariant, and works well with outdoor photos which human face recognition systems have problems with. Torso information is more practical and useful than facial information to recognize people in real world photos.

We use Kullback-Leibler distance measure to classify torsos. However, due to the fact that Kullback-Leibler distance is not symmetric, the sequence of photos affects the result. In addition, the threshold for classification is empirically chosen and may not be optimal for every photo collections. In some cases such as when people wear uniforms, torso based analysis may fail to work properly.

DISSCUSSION AND FUTURE WORK

We plan to perform user studies to evaluate our strategies. Since our implementation is based upon the assumptions on personal image collections, we plan to examine how real usage patterns match our expectation and its implication on hierarchical event clustering and torso based human identification.

This work in semi-automatic image annotation is at an early stage. We have demonstrated an interface that allows users to make bulk annotation effectively by providing appropriate image groups. We suggest that this type of image management tools can help users to manage their personal photo collections easily and enjoyably.

REFERENCES

1. Adobe Photoshop Album, Adobe Systems Inc., http://www.adobe.com/products/photoshopalbum/
2. Bederson, B. B. PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps. Proc. ACM Symposium on User Interface Software and Technology, (UIST 2001), 2001.
3. Cooper, M., Foote, J., Girgensohn, A., Wilcox, L. Temporal Event Clustering for Digital Photo Collections, Proc. of the 11th ACM International Conference on Multimedia, (MM '03), 2003.
4. Duda, R. O., Hart, P. E., and Stork, D. G.: Pattern Classification. 2nd Edition, Wiley & Sons, New York, 2001.
5. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. Query by Image and Video Content: The QBIC System, *IEEE Computer*,28(9), pp.23 -32, Sept. 1995.
6. Gargi, U. Consumer Media Capture: Time-Based Analysis and Event Clustering, Hewlet Packard Tech Report, HPL-2003-165, 2003.
7. Kang, H., and Shneiderman, B. Visualization Methods for Personal Photo Collections: Browsing and Searching in the PhotoFinder, Proc. IEEE International Conference on Multimedia and Expo (ICME2000) New York: IEEE, pp. 1539-1542, 2000.
8. Kuchinsky, A., Pering, C., Creech, M.L., Freeze, D., Serra, B., and Gwizdka, J. "FotoFile: A Consumer Multimedia Organization and Retrieval System", Proc. ACM Conference on Human Factors in Computing Systems (CHI '99), pp. 496-503, 1999.
9. Lienhart, R. and Maydt, J. An Extended Set of Haar-like Features for Rapid Object Detection. IEEE ICIP 2002, Vol. 1, pp. 900-903, Sep. 2002.
10. Nakajima, C., Pontil, M., Heisele, B., Poggio, T., "Full-body Person Recognition System", Pattern Recognition 36(9), pp. 1997-2006, 2003.
11. Phillips, P. J., Grother, P. J., Michaels, R. J., Blackburn, D. M., Tabassi, E., Bone, J. M. "Face recognition vendor test 2002: Evaluation report." NISTIR 6965, 2003.
12. Platt, J.C., Czerwinski, M., Field, B.A. PhotoTOC: Automatic Clustering for Browsing Personal Photographs, Proc. Fourth IEEE Pacific Rim Conference on Multimedia, 2003.
13. Rodden, K. and Wood, K., How do People Manage Their Digital Photographs?, ACM Conference on Human Factors in Computing Systems (CHI 2003), Fort Lauderdale, April 2003.
14. Srihari, R., Zhang, Z., Rao, A., "Image background search: combining object detection techniques with content-based image retrieval (CBIR) systems" IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL '99), 1999.
15. Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M. and Field, B. Semi-Automatic Image Annotation., Proc. Interact '01, pp.326-333, 2001.
16. Yoon, K., Davis, L. Color Path Length, To appear in University of Maryland Tech Report, 2004.
17. Zhao, W., Chellappa, R., Philips, P.J., Rosenfeld, A., Face Recognition: A Literature Survey, ACM Computing Surveys, Vol 35(4), pp. 399-458, 2003.