FotoFile: A Consumer Multimedia Organization and Retrieval System

Allan Kuchinsky, Celine Pering, Michael L. Creech, Dennis Freeze, Bill Serra, Jacek Gwizdka*
Hewlett Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304 USA
+1 650 857 1501
{kuchinsk, celine, dff, creech, bills} @ hpl.hp.com
(* Current address: jacek@ie.utoronto.ca , Interactive Media Laboratory: Department of Mechanical and Industrial Engineering, University of Toronto, 4 Taddle Creek Rd, Toronto, Ontario, Canada M5S 1A4)

ABSTRACT
FotoFile is an experimental system for multimedia organization and retrieval, based upon the design goal of making multimedia content accessible to non-expert users. Search and retrieval are done in terms that are natural to the task. The system blends human and automatic annotation methods. It extends textual search, browsing, and retrieval technologies to support multimedia data types.

Keywords
Multimedia computing, information organization, retrieval, browsing, visualization, content-based indexing and retrieval, digital photography, digital video, metadata, media objects

INTRODUCTION
Technologies and applications for consumer digital media are evolving rapidly. Examples of these technologies are digital still and video cameras, multimedia personal computers, broadband multimedia networks, and recordable CD/DVD. These technologies enable consumers to create and access ever-increasing amounts of content, from a wide variety of sources [1] and formats. As a result, there are significant challenges to be overcome to effectively organize and access this media information.

Consumer research conducted by Hewlett Packard has found that organization and retrieval of digital images is a source of great frustration to customers. Consumers were found to be particularly resistant to the notion of organizing and managing home media, seeing these activities as tedious and error prone. They described photos thrown in shoeboxes and home videos sitting on shelves unviewed.

We derived our approach to making multimedia content accessible to non-experts by

- analyzing the strengths and weaknesses of current commercial products and experimental systems, and
- conducting user research to understand the consumer's perspective on the problem and to gauge customers' reactions to the different approaches.

CURRENT APPROACHES
Technologies for multimedia organization and retrieval have been applied with some success to problems in the business/professional domain. It is not clear, however, that these approaches and technologies are well-suited for consumer-oriented applications. Consumers, in general, have less time, patience, and motivation to learn new technologies.

Traditional keyword-based search technologies are very powerful and flexible. There are a number of commercial image management products that enable a user to search and retrieve visual information based upon indices formed from the user's annotations. Image database products from Extensis (Fetch) [2], Imscape (Kudo Image Browser) [3], Canto (Cumulus) [4], and Digital Now (Showcase) [5] allow a user to browse through files as galleries of thumbnails or as textual lists. The user can typically sort media objects by name, file type, folder, or volume.

The strength of the keyword-based approach is that information about media objects can be expressed in terms that are personally meaningful to the user (i.e., in terms of attributes like creation date, location, subject, and identities of people). Such semantic information about media objects, frequently referred to as *metadata*, provides a rich structure for effective searching. The

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI '99 Pittsburgh PA USA
Copyright ACM 1999 0-201-48559-1/99/05...$5.00

IPR2025-01314; Samsung
Ex. 1021 - Page 1

disadvantage is that making such metadata available usually means that keywords and textual annotations must be entered manually. This works for business applications, where there is an economic incentive for time and effort being devoted to indexing activities. Lacking these same economic incentives, consumers are more resistant to the task of data entry.

An alternative approach, content-based indexing and retrieval, provides some degree of automation for this process by automatically extracting features, such as color or texture, directly from visual data [6]. Products from Virage [7] and IBM (QBIC) [8] implement mechanisms for content-based retrieval of images. By using the intrinsic visual attributes of images, such as color, structure, texture, and composition, to perform queries; users can search collections by instructing the system to retrieve images that are visually similar to the sample image. Images returned by the query are ordered by the degree of similarity to the base image.

The content-based indexing and retrieval approach frees the user from the task of data entry, and it utilizes people's perceptual abilities. These technologies work well in situations where a user wants to locate a visual image that is similar to a sample image. The disadvantage is that these systems only extract low-level syntactic features (measures of color and texture), which are not as personally meaningful to consumers as keyword-based attributes.

An additional concern we had with current technological approaches was whether they correctly map to consumers' likely information-seeking behaviors. Much attention has been paid to the task of direct search, in which a user knows the target. Relatively little attention has been paid to the activities of browsing through collections of materials, where the user doesn't have a very specific goal in mind, and serendipitous discovery is important [9]. It is likely that browsing will be a preferred information seeking behavior for consumers, and it should, accordingly, receive more systematic support from search/discovery technologies.

USER RESEARCH

To understand the consumer's perspective on the multimedia organization and retrieval problem and to gauge customers' reactions to the different approaches, we conducted a set of focus group sessions in the Denver and San Francisco areas. We were looking to more fully understand how people inherently organize visual materials and, in particular, to gather information on the perceived tradeoffs between

- manual vs. automated annotation, and
- direct search vs. browsing.

In order to understand the differences between business and consumer usage, we held different focus groups for business and home participants, respectively.

The sessions began with a discussion of how the participants currently organize, find, and share photos. This was followed by a group exercise in organizing a set of travel photos. We then presented participants with mockups of concepts for keyword-based indexing and search, visually-based search (content-based indexing and retrieval), and visual overview (browsing).

Our key findings from these sessions were that:

- Keyword-based search was the easiest concept for home participants to grasp. However, they saw drawbacks, both in the time-intensive nature of entering keywords for photos and in the possibilities for many false "hits" while searching.
- Participants readily grasped the benefits of automated indexing. However, the home participants thought that they would use keyword-based search more frequently.
- Home participants reacted very favorably to the notion of browsing, much more favorably than did business participants.

We drew two conclusions from these findings; first, that consumers would desire the benefits of both keyword-based search and automated indexing; second, that there may be a considerable role for browsing techniques in supporting consumers' multimedia information seeking activities.

THE *FOTOFILE* SYSTEM

Based upon our analysis of current approaches and our findings from user research, we developed a hybrid approach to address the problems of multimedia organization and retrieval for consumers. We prototyped a number of techniques which make it easier for consumers to manually annotate content and to fit the annotation task more naturally into the flow of activities that consumers find enjoyable. We also utilized a number of automated content-based indexing techniques in order to both substitute for manual annotation where appropriate and to provide novel capabilities for content creation and organization. Finally, we augmented direct search tools with techniques for browsing and visualization of large digital media collections.

FotoFile, shown in Figure 1, is an application for organizing and managing consumer digital media, such as photos and audio/video recordings. It illustrates a number of aspects of our hybrid approach.

FotoFile displays multimedia in a photo-centric way by displaying *media objects* that consist of a photo with related sound and video attached. For video content,

IPR2025-01314; Samsung
Ex. 1021 - Page 2

FotoFile generates photos by extracting keyframes from the video. The leftmost pane is a *Content Index*, which enables the user to annotate and search for materials. An *a priori* set of pre-defined metadata attributes is used to represent common properties of media objects, such as *creation date*, *location*, *subject*, *people*, *title*, and *description*. Users can assign arbitrary values within the defined *metadata* types, e.g. annotating the *location* of a photo as "Grand Canyon". Another pre-defined metadata attribute, called *favorite*, can be used to tag certain images as the "best" images in a collection, e.g. my favorite photos from the Grand Canyon vacation.

The central pane is an *Image Palette*, which provides functionality analogous to a light table. The user can arrange, delete, and display media objects at different resolutions in the *Image Palette*. The palette is also used to display search results and newly imported materials, and it also serves as a temporary storage area for creating albums.

The rightmost pane is an *Album Editor* that provides tools for composition of digital albums, which can then be "played back" or sent electronically to others.

In order to match the user's expectations for how pictures are arranged, *FotoFile* uses a photo album as the primary organizational metaphor. A photo album is a metaphor with which people can quickly relate when thinking about organizing photos, and therefore the mental model relies on user intuition rather than explicit instruction. In *FotoFile*, an *Album* is a persistent collection of media objects, which are arranged on "pages". Each image is also accompanied by annotations, which can be in the form of text, audio, or video. Furthermore, in order to simplify album retrieval, the user can assign a representative image for the album cover to aid in selection from a list. Having a cover image that is representative of the album in the user's mind enables fast visual recognition, rather than relying on information recall.

Techniques to Ease the Task of Manual Annotation

Bulk Annotation

We provide mechanisms for bulk annotation, which enable the user to quickly annotate large numbers of items with a minimal number of gestures. For example, the user can select multiple media objects in the *Image Palette*, select several values within the *Content Index*, and then press the Annotate button. This results in the assignment of all selected values to all of the selected media objects.

Symmetry between Annotation and Search

Since we were designing *FotoFile* for home usage, we designed the annotation and search interfaces to use the same basic mechanism. There is a visual and gestural symmetry between the actions for annotation and search. Users only need to learn one tool for both activities.

To annotate content, the user selects one or more metadata attribute/value pairs, and presses the Annotate button. At that time, the selected attributes are applied to all selected media objects. To retrieve content, the user again selects one or more attribute/value pairs, and presses the Search button. At that time, all media objects that have the selected attributes are immediately displayed in the *Image Palette*. There are several search modes, including Boolean operations and a similarity-based search built upon automated feature extraction [18].

Since there is no default mode, the user is free to intermix the annotation and search activities, which we believe will result in a better-annotated corpus of material than would occur if the user only had a dedicated authoring mode available.

Use of Narrative Structure to Help Organize Content

Annotating content manually is time consuming, and it transforms the process of creating photo albums from an enjoyable activity into a very tedious one. On the other hand, people like to tell stories with photos [10] and the organization of photos into stories can provide us with a significant amount of information that can serve as metadata. That is, we can use narrative structure underlying the events captured in photos as a source of their organization and annotation. This effectively turns the organization process into a storytelling activity, an activity that is more enjoyable than the task of organization, which carries with it the connotation of "work".

Whereas with conventional photography, storytelling is typically done using prepared albums and collages, whose structures are fixed, digital photography allows the user to employ more dynamic collections of photos in storytelling. The user can arrange small groups of photos into segments that correspond to single narrative episodes. These segments can be reused in different situations and combined in different ways, depending upon the interaction between storyteller and audience. The model of usage is of two or more people sitting together by a computer, much in the same way that people sit together and go through photo albums. An alternative model of usage is one wherein the storyteller shares groupings of photos and annotations over the Internet.

IPR2025-01314; Samsung
Ex. 1021 - Page 3

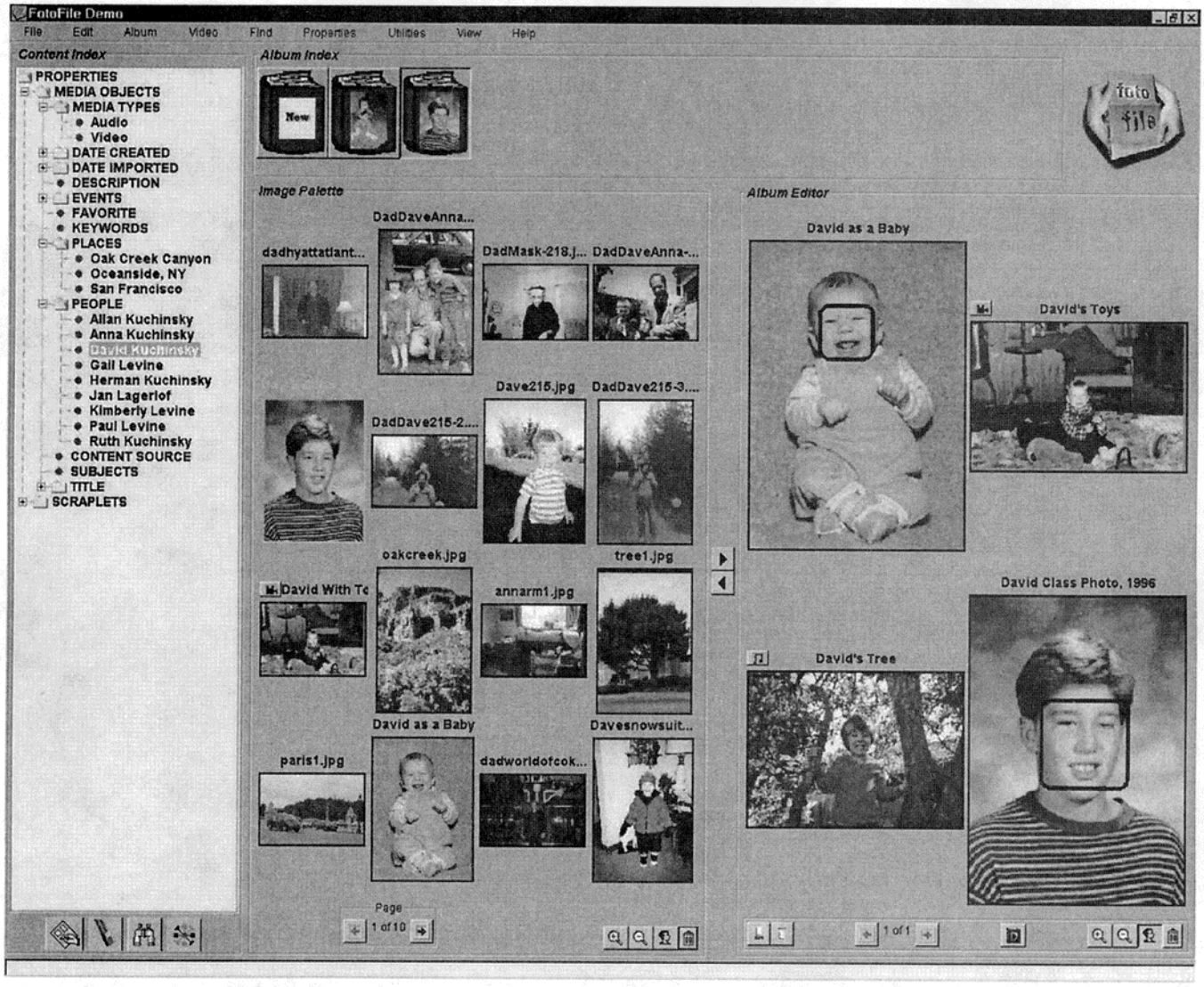

Figure 1. Building a Multimedia *Album* in *FotoFile*.

Building on the metaphor of a scrapbook, we call these small groupings of photos scraplets (shown in Figure 2). A scraplet can be assigned a name and other properties, thus providing annotation for a grouping that can be useful in retrieving the grouping at a later time. We believe that such grouping and lightweight annotating will fit naturally within the activity of preparing a story, thus providing a more enjoyable mechanism for eliciting metadata from consumers. Moreover, use of voice annotation may bring additional emotional power to stories that are shared over the Internet.

The selection of photos for grouping into scraplets is based upon two assumptions. *First,* the user should have a personal memory of the events depicted in the photos. *Second,* chronological ordering of events is a dominant organization principle of human episodic memory [11]. Using the same photos in multiple scraplets links them implicitly. The links are displayed during album playback to indicate to the user multiple possible story lines.

Benefits of Automated Feature Extraction

The use of automated feature extraction tools enables *FotoFile* to generate some of the annotation that would otherwise have to be manually entered. It also provides novel capabilities for content creation and organization.

Face *Recognition*

The black rectangular highlights on the pictures of David in Figure 1 denote faces that have been recognized by a face detection and recognition system [12] [13]. Information about recognized faces appears in the *Content Index* in an identical manner to metadata gathered by human annotation. This is one example of the integration of automated and human annotation in our approach, and it results in a hybrid system where the user guides the mechanisms.

When given photos that contain faces of new people, the face recognition system attempts to match the identity of the face (see Figure 3). The user either corrects or confirms the choice; the system then can more accurately

match faces to their correct identities in subsequent photos. Once a face is matched to a name, that name will be assigned as an annotation to all subsequently seen photos that contain faces that match the original. To handle the false positives and false negatives of the face recognition system, a user must confirm face matches (see Figure 4) before the annotations associated with these faces are validated (i.e., added to the *Content Index).* Users view the matched identities of faces through tooltips displayed when the mouse sprite enters the: rectangular highlight surrounding a face.

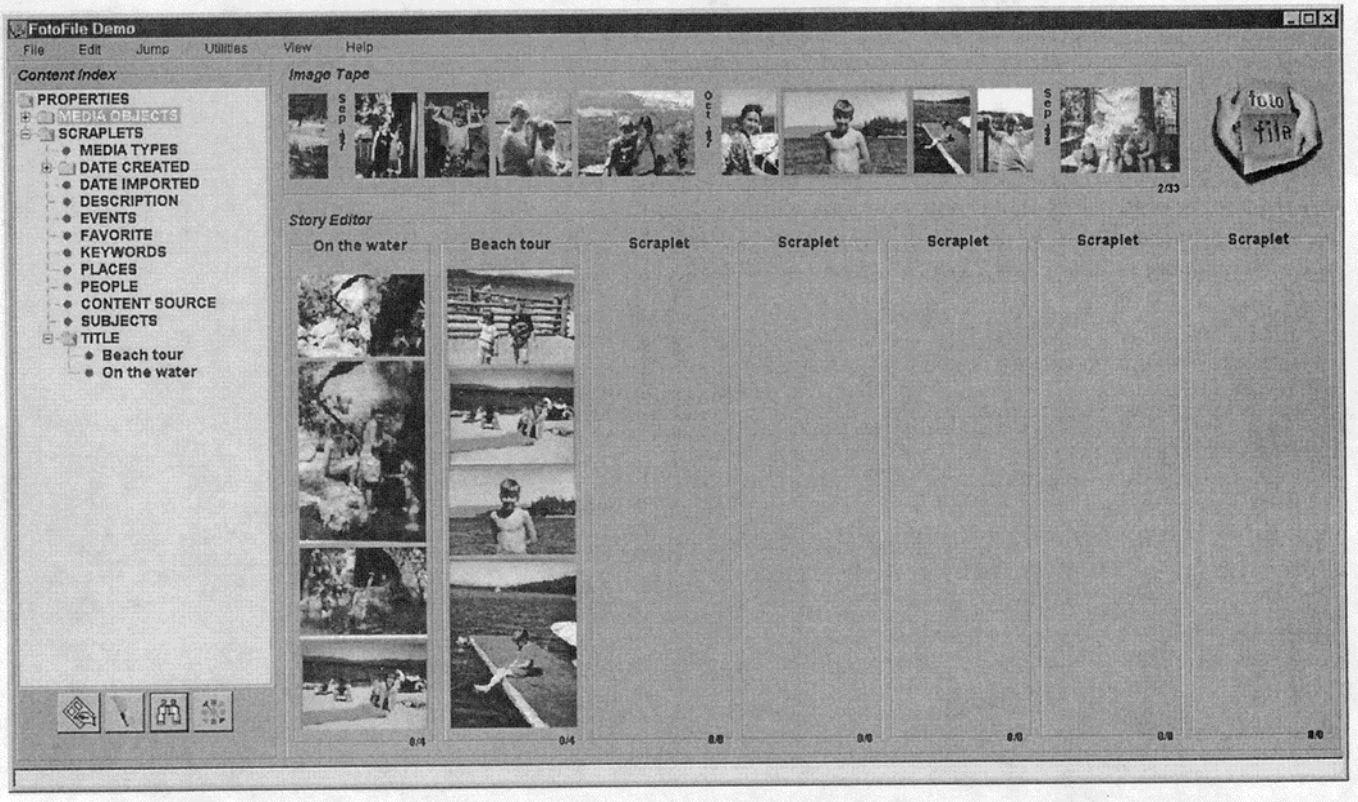

Figure 2. *Scraplets* created in the Story-Editing Environment. Organizing photos via multimedia "scraplets" reduces the tedious effort of manual annotation.

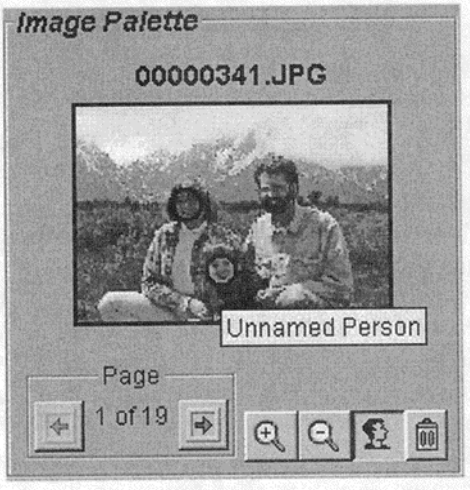

Figure 3. First photo of Merrick is not matched to any other faces by the recognizer; user enters name *Merrick*.

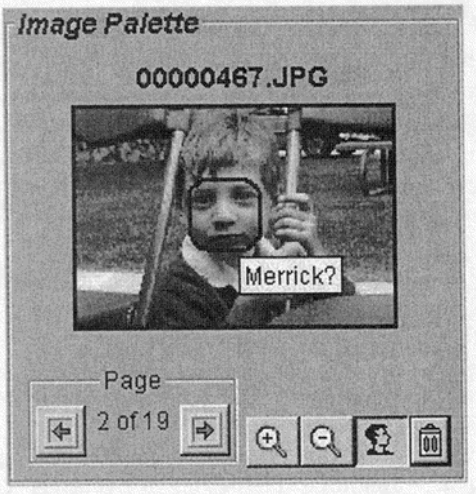

Figure 4. Subsequent photo of Merrick is matched by the recognizer.

IPR2025-01314; Samsung
Ex. 1021 - Page 5

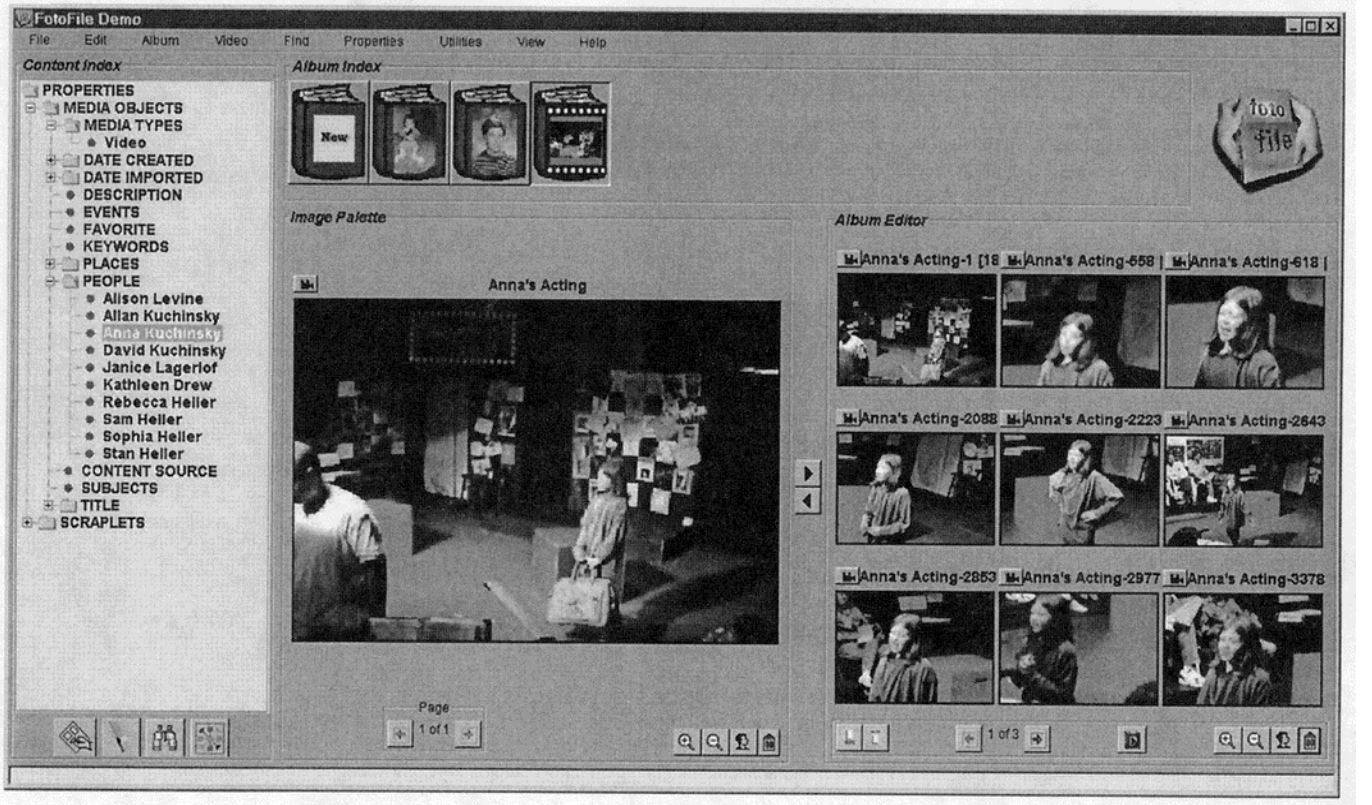

Figure 5. An Automatically Generated *Video Album.*

Video Shot Detection

The ***FotoFile*** user can automatically generate "albums" of video clips extracted from longer video sequences using the video shot detection and keyframe extraction algorithms [14][15]. Video shot detection is the process of detecting boundaries between consecutive shots so that sequences of interrelated video frames can be grouped together. Examples of shot boundaries include abrupt shot changes caused by turning the camera off, as well as more sophisticated shot transitions like fades, dissolves, and wipes. A user can easily create an album that contains a series of video clips that comprise a video (see Figure 5). Each clip represents a playable segment of video. Since each video segment is itself a media object, it can be rearranged, or placed in different albums-just like any ***FotoFile*** media object.

Video Keyframe Extraction

During the shot detection process, a keyframe extraction algorithm [15] is used to generate a set of video frames (still images) which best represent the content of each shot. These keyframes attempt to represent abrupt changes in video content as well as slower, ordered changes like pans and zooms. Each resulting keyframe is associated with a video clip that starts with that frame and continues to the end of the shot. The set of these keyframes imposes an extra structure on shots which help users fine-tune their selection and manipulation of video clips and shots.

Keyframe extraction is also used to derive a representative picture for each video imported into ***FotoFile.***

Browsing and Visualization of the Content Space

We believe that consumers' information-seeking activities differ from those of information retrieval professionals, and that this is particularly true when the information involves home media such as photos or videos.

In these settings, directed search may be less frequent, whereas riffling and browsing through collections of materials becomes the norm (and serendipity is expected).

We provide support for these activities by integrating visualization and browsing tools into ***FotoFile,*** such as the ***Hyperbolic Tree*** package from Inxight Software [16][17]. Figure 6 shows a hyperbolic tree built from the attributes and values in the ***Content Index.***

One problem observed in usability studies of the ***Hyperbolic Tree*** [21] was that items on the outside rim of the display tended to group strongly, with users often assuming that they belonged in the same category. It was suggested that careful use of alternative perceptual coding for semantic categories could alleviate this problem. We have achieved this by providing additional views based on

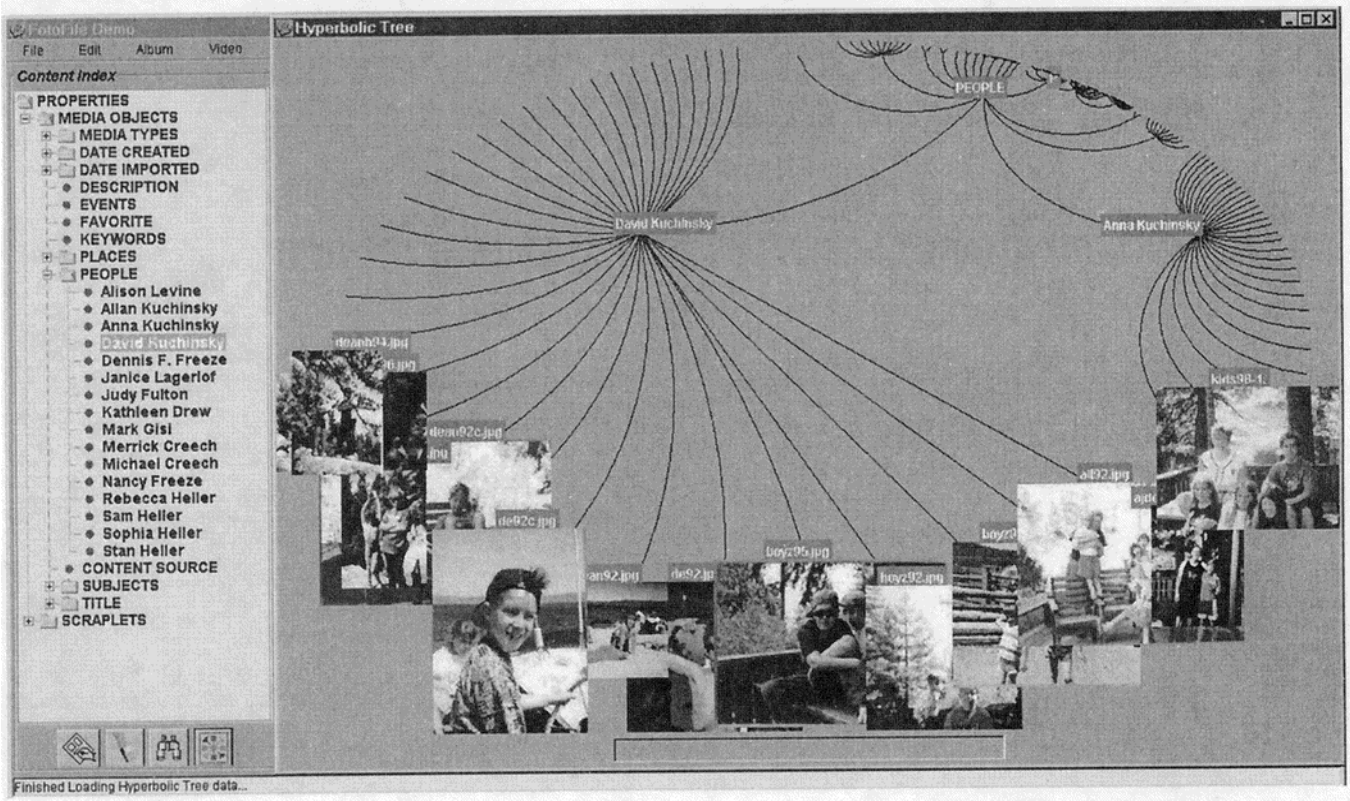

Figure 6. Viewing the *Content Index* via the *Hyperbolic Tree.*

the use of automated image feature-extraction software [18]. Image content is analyzed to extract measures for color distribution and texture, and a clustering algorithm [19] recursively partitions the collection of media objects to form the tree model displayed by the *Hyperbolic Tree.* In this way, media objects that are visually similar to each other will appear closer to each other in the visualization space. This adds structure to the browsing activity, enabling the user to visualize related clusters of materials in an intuitive manner.

DISCUSSION

With *FotoFile,* we have attempted to balance tradeoffs across two dimensions of information-seeking behaviors:

- Combining the strengths of both human annotation and automated feature extraction.
- Accommodating both directed search and exploratory browsing and visualization.

Based upon our findings from user research, we have attempted to integrate these capabilities in a way that is suited to the needs of the consumer environment. In order to provide an integration that is easily understandable and usable, we need to emphasize certain capabilities more than others. To determine the appropriate balance points, additional user research is needed. In particular, we need to determine:

- The degree to which consumers will perform annotation if the benefits are significant and meaningful.
- The usability and usefulness to consumers of browsing and visualization environments.

One challenge in designing credible studies of this nature is in defining the right metrics for data analysis. Consumer information-seeking behavior is different from that of specialists performing directed searches in textual databases, where large numbers of people are searching over large information spaces for materials indexed by some unknown person. The characteristics of an information-seeking environment for consumers involve relatively few people searching (e.g., immediate family members) over a small amount of information (less than several thousand items in a collection) that they have personally indexed, or that was indexed by someone they know. In many cases, serendipitous discovery is a significant (but often unstated) goal. The traditional metrics of *recall* and *precision* may not be as applicable. Alternative measures might include the level of goal attainment, the efficiency (number of actions) to reach a goal, the utility of the information found, which annotations and features are used for later retrieval by both novices and experts, and subjective measures of user satisfaction [20].

IPR2025-01314; Samsung
Ex. 1021 - Page 7

CONCLUSION

We have built an experimental multimedia organization and retrieval system that attempts to balance tradeoffs between (1) human annotation versus automated feature extraction, and (2) directed search versus exploratory browsing and visualization. The ultimate goal is to make multimedia content accessible to non-expert users.

Photography and home movies are activities that address deep human needs; the need for creative expression, the need to preserve memories, the need to build personal relationships with others. Digital photography and digital video can provide powerful and novel ways for people to express, preserve, and connect. However, new technologies often raise new problems; the problem of multimedia organization and retrieval is brought about by the very technology that makes it possible for people to create and access ever-increasing amounts of content, from a widening diversity of sources.

By helping consumers to better manage content, we hope to enable people to take full advantage of the benefits provided by digital media technologies.

ACKNOWLEDGMENTS

We owe a great debt to HongJiang Zhang, John Wang, and Wei-Ying Ma for their excellent work in content-based indexing and retrieval, which has been incorporated into our prototypes. Thanks and praise to Rick Steffens, Mike Krause, and others at HP's Colorado Memory Systems Division for their support, encouragement, and inspiration. Ella Tallyn made substantial contributions to both the visual design of *FotoFile* and the conceptual design of our use of narrative structures in *FotoFile*.

REFERENCES

1. Kuchinsky, A., Bit Velocity is Not Enough: Content and Service Issues for Broadband Residential Information Services, *IEEE 3rd International Workshop on Community Networking*, Antwerp, Belgium, May, 1996.
2. Extensis Corporation, http://www.extensis.com/.
3. Imspace Systems Corporation, http://imspace.com/.
4. Canto Software, http://www.canto-software.com/.
5. Digital Now, http://www.digitalnow.com/.
6. Furht, B. Smoliar, S., Zhang, H., and Furht, B. *Video and Image Processing in Multimedia Systems*, Kluwer Academic Publishers, 1995. Conger., S., and Loch, K.D. (eds.).
7. Virage Incorporated, http://www.virage.com/.
8. IBM, http://www-i.almaden.ibm.com/cs/showtell/qbic/.
9. Chang, S.J., and Rice, R.E., Browsing: a Multidimensional Framework, in Williams, M.E. (ed), *Annual Review of Information Science and Technology*, Vol. 28, pp. 231-276, Medford, NJ, 1993..
10. Chalfen, R. *Snapshot Versions of Life*. Bowling Green State University Press, Bowling Green, Ohio, 1987.
11. Tulving, E. *Elements of Episodic Memory*. Oxford, UK: Oxford University Press, 1983.
12. Turk, M., and Pentland, A. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71-86, 1991.
13. H.A. Rowley, S. Baluja and T. Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23-38, Jan. 1998.
14. H.J. Zhang, C. Y. Low and S. W. Smoliar. Video parsing and browsing using compressed data. *Multimedia Tools and Applications*, vol. 1, pp. 89-111, 1995.
15. H.J. Zhang, et al. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, Pergomon Press/Pattern Recognition Society, May 1997.
16. John Lamping, Ramana Rao, and Peter Pirolli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (May 1995), ACM.
17. Inxight Software, Inc., http://www.inxight.com.
18. W.Y. Ma and H.J. Zhang. Content-based image indexing and retrieval. Chapter 13, *The Handbook of Multimedia Computing*, edited by Borko Furht, CRC Press LLC, 1998.
19. R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley Publications: NY, 1973.
20. Wilson, K. Evaluating Information Exploration Interfaces, position paper for Workshop on Innovation in Information Exploration Environments, *CHI'98 Conference on Human Factors in Computing Systems*, http://www.fxpal.com/CHI98IE/.
21. Czerwinski, M. and Larson, K., Trends in Future Web Designs: What's Next for the HCI Professional?, *ACM Interactions*, November-December, 1998, pp. 9-14.