Multimedia Information Retrieval at FX Palo Alto Laboratory

Matthew Cooper, John Adcock, Andreas Girgensohn, Jeremy Pickens, and Lynn Wilcox

FX Palo Alto Laboratory, Palo Alto, CA 94034 USA

ABSTRACT

This paper describes research activities at FX Palo Alto Laboratory (FXPAL) in the area of multimedia browsing, search, and retrieval. We first consider interfaces for organization and management of personal photo collections. We then survey our work on interactive video search and retrieval. Throughout we discuss the evolution of both the research challenges in these areas and our proposed solutions.

1. INTRODUCTION

The *Multimedia Access and Visualization* (MAV) research group at FXPAL conducts projects spanning multimedia processing and information retrieval. In this paper, we review systems built over the last several years to address two core multimedia problems: digital photo organization and interactive video search. We document the evolution of our work through a period of substantial changes in computing capabilities and consumer practices. In particular, visual content is now captured and distributed at unprecedented scale. As a result, the demand has increased for tools that leverage automatic processing to aid in content management. Our objective has been to design systems using automatic analysis to minimize required manual effort, while maximizing its utility with flexible interaction techniques and visualizations.

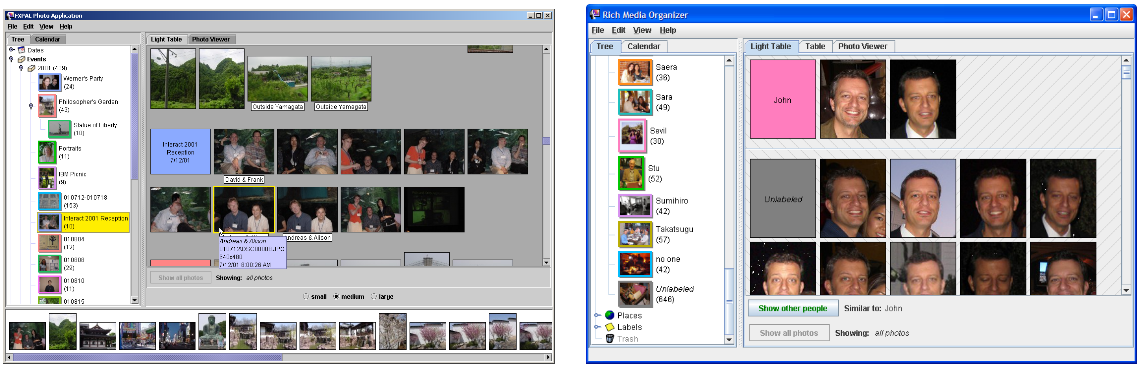


Figure 1. The FXPAL Rich Media Organizer. Left: light table view. Right: interactive face labeling.

2. DIGITAL PHOTO ORGANIZATION

A recurring research focus has been the management of personal photo collections. At FXPAL work in this area originated with the Rich Media Organizer (RMO),[1] whose design concentrated on enhancing the user experience without neglecting the mundane components of a photo organization application. Hardware advances continue to reshape consumer digital photography practices, and the distribution of photos within a variety of social networks is widespread. More recent systems at FXPAL enable photo grouping and selection by leveraging rich metadata now recorded at the time of capture.

Further information: http://www.fxpal.com

2.1 FXPAL Rich Media Organizer

The Rich Media Organizer was targeted for a typical desktop PC to enable browsing, tagging, and organization of potentially large personal photo collections. To facilitate navigation, the RMO automatically divides photos into meaningful episodes or events,[2] such as a birthday party or a trip. This automatic segmentation of the collection is based solely on the photo time stamps. The RMO presents the user's entire photo collection in a scrollable light table of thumbnails as on the left of Figure 1. Markers indicate the start of each event. A tree view for events and other attributes (people, places, etc.) can scroll the light table to a selected event, sort photos by different categories, or filter the set of visible photos to show only a particular category. Anticipating advances in current photo organizers, the RMO also integrated face detection[3] and semi-automatic interactive person tagging (right panel of Figure 1),[4] collage creation,[5] and automatic slideshow creation.

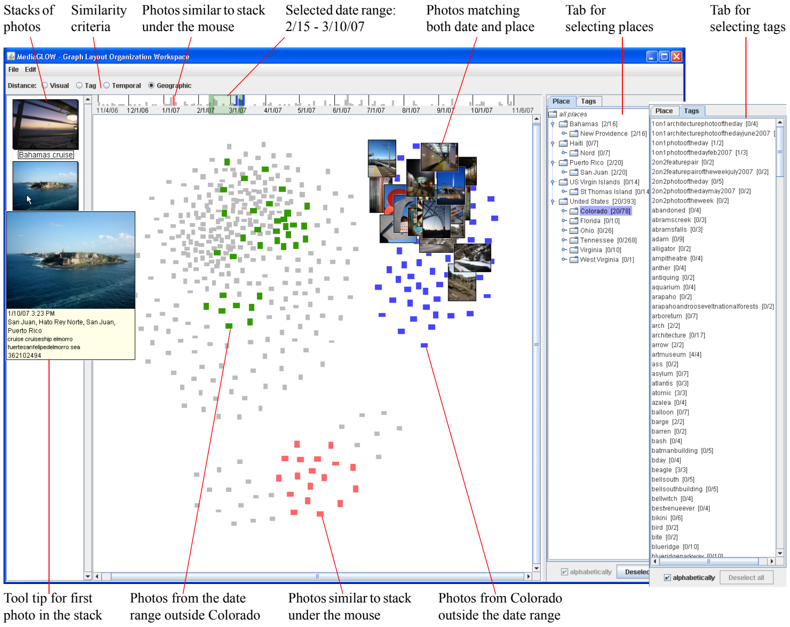


Figure 2. A view of MediaGLOW.

2.2 MediaGLOW

MediaGLOW explores the organizational benefits of an interactive visual workspace using multiple similarity criteria to layout photos.[6,7] Today, it is common for users to accumulate large photo collections. Locating photos of interest in such collections requires multiple photo attributes. For example, when looking for photos of sunsets from Italy, one could filter by geographic location and then use a layout by visual similarity to place photos with shades of orange near each other. One could also use the time of the day to select among the Italy photos, assuming that the camera was set to the correct time zone. It would be even easier if those photos were tagged with "sunset."

MediaGLOW offers offers four different similarity criteria: temporal, geographic, tag, and visual. Tag similarity is computed using the Jaccard similarity coefficient of tags shared across photos. Our visual similarity is determined by image classifiers trained on manually tagged photos from Flickr that compares predicted likelihoods for tags. Besides grouping photos by similarity, MediaGLOW also provides filters that restrict the time range, the geographic location, and the tags assigned to matching photos.

MediaGLOW integrates a variety of visualization and interaction techniques to enable users to find photos by proximity and by attribute filters. It uses a graph layout mechanism to visually indicate similarity among photos in the space while optimizing desired distances between photos. While grid-based layouts are more common for photo applications, they cannot accurately present similarity by proximity for all of our criteria of interest.

MediaGLOW computes a layout once for each similarity criterion and treats each as a separate layer. Users can rearrange photos in each of those layers. When revisiting a specific layout, photos remain where the user left them.

Figure 2 shows a screen shot of the MediaGLOW interface. Users group photos into stacks as shown in the sidebar to the left of the workspace. Stacks enable the creation of and access to user-definable categories. In addition, stacks are used by the system as examples for identifying photos similar to those already in the stacks. Users may hide photos that do not match particular criteria. Time can be restricted with sliders in the timeline (see the top of Figure 2). Geographic location and user-assigned tags are specified in tabs (see the right of Figure 2). A list of all tags assigned to photos can filter photos by selected tags. Similarly, selecting a location makes only the corresponding photos visible. When hovering with the mouse over a stack, similar photos per the selected criteria are highlighted in the workspace and timeline (see pink dots and timeline bars in Figure 2). This provides the user with guidance for further exploration. A more complete description of MediaGLOW appears in,[7] which includes a user study that demonstrates the benefits of combining multiple search attributes in a common interface.

The progression from the RMO to MediaGLOW reflects changes in consumer photography practices. The tools demanded by consumer digital photographers have moved beyond simply organizing their collections towards repurposing and sharing their content. The RMO remains a powerful tool for browsing and structuring individuals' photo collections. MediaGLOW supports users in "collect then select" activity on which content repurposing and sharing relies. For example, users may wish to design on-line slideshows or printed photo albums by sifting through accumulated photos from a specific event, place, or time. Consumer photography continues to shift towards mobile networked devices that record rich metadata at the time of capture. Heterogeneous personal media collections comprised of video and still images captured using mobile phones present new opportunities for content management and repurposing tools.

3. INTERACTIVE VIDEO RETRIEVAL

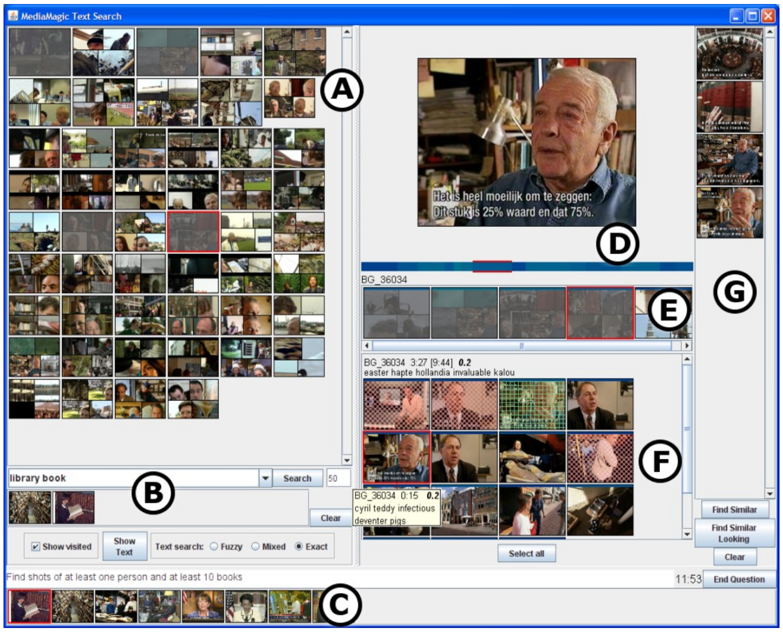
Our work on video indexing, browsing, and retrieval can be traced back to the MBase system[8] which focused on facilitating access to videos captured and archived at FXPAL. MBase's emphasis was on providing flexible access to slide-based presentations captured in (what were at the time) heavily instrumented conference rooms. MBase featured a powerful video player, novel keyframe-based interfaces into retrieved videos,[9] and a web-based client.

Later, we participated in the TRECVID retrieval evaluations,[10] which allowed us to benchmark our interfaces and retrieval systems against competing approaches. We also scaled our analysis methods and systems moving towards the present era of more abundant and more freely distributed video resources. Recently, our work has come full circle. Our efforts have shifted again towards slide-based presentation videos, but now at internet scale. The TalkMiner[11] system aggregates lecture webcasts across multiple web sites, and performs analysis and indexing to make the content more effectively and efficiently searchable. In the next section, we review systems built through this progression.

3.1 MediaMagic

Our group developed an interactive video search system, MediaMagic,[12–14] to enable users to efficiently search video using a flexible interface and rich visualizations. MediaMagic provides tools for issuing queries using textual, visual, and semantic content. The system interface appears in Figure 3. Text or image search results are displayed in query-dependent summary visualizations that encapsulate the relationship between the query and the search result in multiple visual dimensions. As the user steps into their search results, these cues are maintained in the representations of keyframes and timelines, along with visual cues that indicate navigation history and existing relevance judgments. Search by textual or visual example is enabled throughout to aid further exploration. Indexing at both shot and story levels is used to create search indices based on available automatic speech recognition generated text transcripts.

We applied variations of the MediaMagic system to the TRECVID interactive search evaluations from 2004-2008. While maintaining a high level of performance, we consistently approached TRECVID as an opportunity



library book

Find shots of at least one person and at least 10 books

11:53

Het is heel moeilijk om te zeggen:
Dit stuk is 25% waard en dat 75%.

BG_36034

BG_36034 3:27 [9:44] ***0.2***
easter hapte hollandia invaluable kalou

BG_36034 0:15 ***0.2***
cyril teddy infectious
deventer pigs

Figure 3. Interactive video search interface. (A) Search results area with story keyframe summaries. (B) Search text and image entry. (C) Search topic and example media. (D) Media player and keyframe zoom. (E) Story timeline. (F) Shot keyframes. (G) List of relevant shots.

to assess novel research directions in visual information retrieval. We next describe several interactive video search systems we developed in these efforts.

3.2 Text-free multimedia search

We implemented a version of our system that makes no use of text during pre-processing or search by altering the previously described MediaMagic system. To determine a story-level segmentation we use semantic concept vectors[15] to represent each shot instead of the text transcript. We use a novelty-based segmentation[16] to locate story boundaries and preserve the multi-level indexing structure used in MediaMagic. Next, we disabled the text query box and the text-based similarity searching. Query relevance in the text-free system is determined solely by similarity between extracted color correlograms (with the “find similar looking” and image query operations), and semantic concept vectors (with the “find similar” operation). An interesting lesson from the evaluation in 2007 was that removing the capacity for text search and substituting our semantic similarity during story segmentation did not significantly degrade performance.[17] We argue that the indexing utility of our automatic semantic analysis is comparable with that of the provided cascaded output of automatic speech recognition and machine translation.

3.3 Collaborative video retrieval

Intuitively, interactive information seeking can be performed more effectively as a collaboration than as a solitary activity.[18] Collaboration between multiple searchers represents a means to further advance search systems that complements improvements in automatic content analysis. We incorporated MediaMagic into a two-user collaborative exploratory search system which comprises a set of interfaces and displays, a middleware layer for handling traffic, and an algorithmic engine.[19] By interacting through system-mediated information displays, searchers help each other find relevant information more efficiently and effectively than they would working alone. Each searcher on a team may fill a unique role, with appropriately customized interface and display components.

in which to evaluate this idea. At TRECVID, our first collaborative system augmented MediaMagic with a rapid serial visualization (RSVP) result browsing interface.[20] The RSVP interface, shown in the left panel of Figure 4, is designed for relevance assessment of video shots, which are presented in a rapid but controllable sequence. A shared display showed continually-updating information about issued queries, all shots marked as

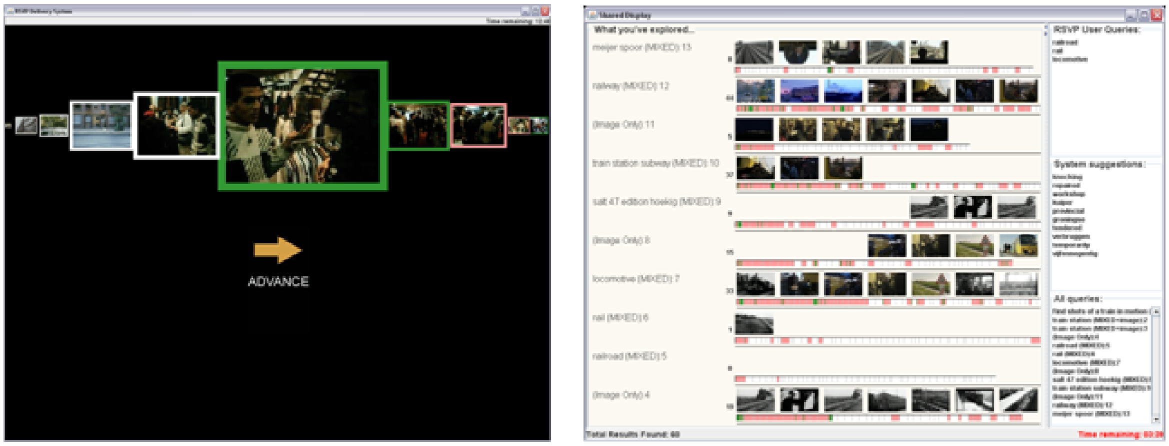


Figure 4. Left: RSVP interface. Right: shared display interface.

relevant by either user, and system-suggested query terms based on activities of both users.[21] The shared display was easily viewed by both users, and appears in the right panel of Figure 4.

Evaluation results indicated a consistent advantage for collaborative search over merged results from two independent single-user searches.[17] The relative effectiveness of collaborative search was in part determined by the number of available relevant shots in the corpus for a given search topic.[19] When there are relatively many relevant shots to be found, it appears that searchers in a time-limited task should be freer to work on their own, without algorithmic collaboration. There are enough available relevant shots that each independent searcher can spend their allotted time working separately. However when relevant information was scarce, two searchers were able push further into the collection and locate more relevant information only with algorithmically collaborated search.

3.4 Cluster-oriented Retrieval

Relevant shots tend to cluster within a small subset of the programs. In the 2007 TRECVID dataset, 50% recall (on average across all topics) is achievable with shots from only the three best programs. We thus added a "Program Mode" view, shown in the left panel of Figure 5 to our system. A keyframe/video viewer is available at the top, and shots in the current program are shown in time-order underneath. Below is a visualization of which programs have yielded the most relevant results (blue bars) or the cumulative score the program has received (yellow bars). At the bottom, ellipses indicate programs with stories which have been frequently retrieved, but not yet examined. Shots judged relevant are arranged vertically on the right. This mode is accessed at the searcher's discretion through a tab on the main interface pane and provides a streamlined interface for reviewing all the shots within a single program or video file.

The aim is to find those programs with the highest available number of relevant shots. Before the search session ends, the user can revisit those programs with the highest scores and search through them in greater detail. The shared display was altered as in the right panel of Figure 5. The blue component at the top indicates which programs contained the most relevant shots. The yellow component shown in the middle indicates programs with stories which are repeatedly retrieved but not yet viewed. The ellipses indicate programs with the highest cumulative retrieval score aggregated across all searches. The text boxes at the bottom show the most frequent terms from the aforementioned different program categories. In 2008, the collaborative system achieved approximately 30% higher mean average precision (MAP) than the best single user run, and 100% higher MAP than our single-user baseline system from 2007.

4. TALKMINER: A LECTURE WEBCAST SEARCH ENGINE

During the period we participated at TRECVID, video content was proliferating rapidly on the internet. While much of this growth has been in short-form, user-generated content, lecture webcasts also represented a steadily

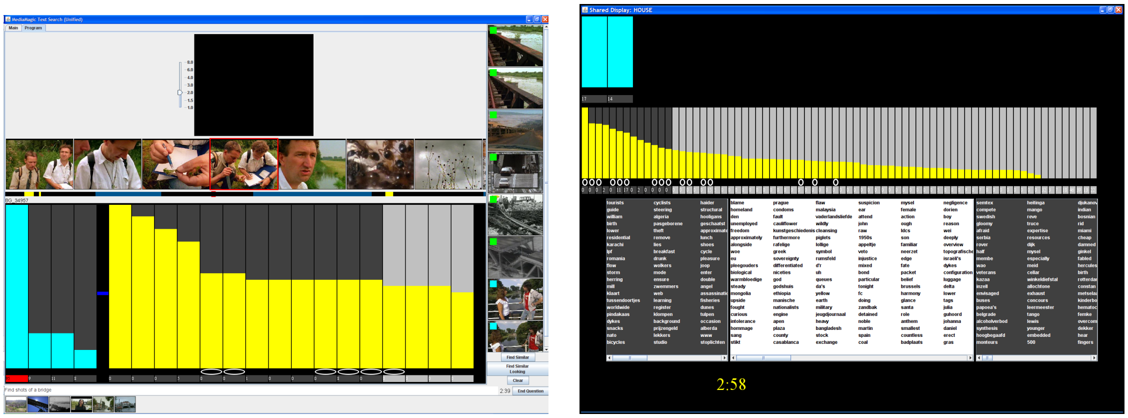
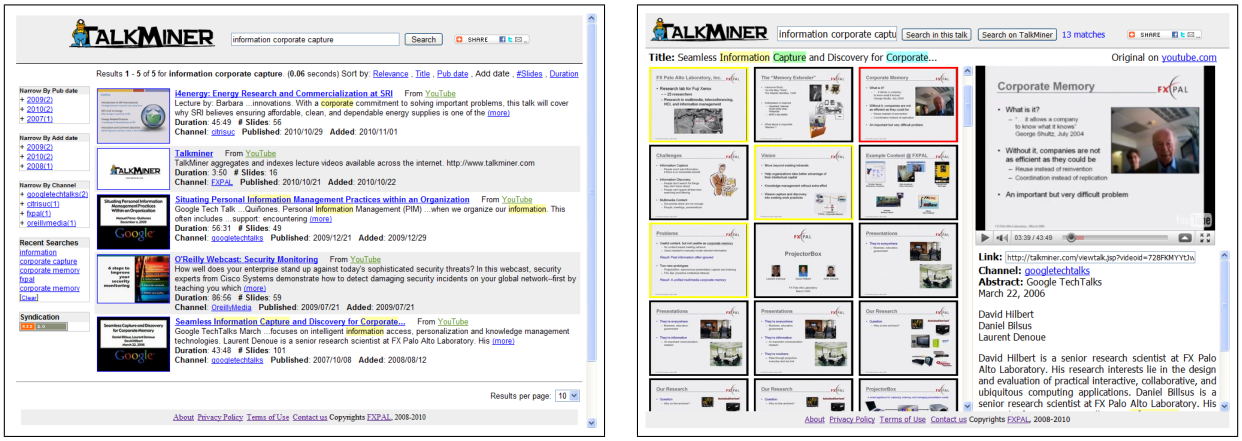


Figure 5. Left: the program mode interface. Right: the shared display interface.

growing segment. These webcasts may be class lectures (e.g., Berkeley Webcast, MIT Open Courseware, etc.), research seminars (e.g., Google Tech Talks, PARC Forums, etc.), product demonstrations, or training materials. Conventional web search engines can often locate these lecture videos, but only when supporting text appears on the hosting web page, or the media has been tagged or otherwise authored within a purposed hosting system. But users, especially students, need to find the locations within a video when a speaker discusses a specific topic. Addressing this need requires a search engine that can identify relevant material *within* the webcast. The TalkMiner system builds on FXPAL's history in both meeting capture and indexing, and our experience searching generic, heterogeneous content at TRECVID.

TalkMiner builds its search index and interface from commonly recorded video. It requires neither dedicated lecture-capture systems, nor careful post-capture authoring, nor even constraints on the style of the video capture. TalkMiner leverages existing online video distribution infrastructure to embed webcasts within an enhanced interface.[11] This approach minimizes storage and bandwidth requirements for scalability and portability. Thus, the system can scale to a greater volume and variety of existing and newly created content at a much lower cost than would otherwise be possible. The system comprises two main components: the back-end video indexer and the front-end web server. The video indexer assembles lecture webcasts by parsing RSS feeds to which the system is subscribed. TalkMiner automatically identifies slide images within each downloaded video and processes them via optical character recognition (OCR) to create its text search index.



information corporate capture

Results 1 - 5 of 5 for information corporate capture. (0.06 seconds) Sort by: Relevance , Title , Pub date , Add date , #Slides , Duration

Narrow By Pub date
+ 2009(2)
+ 2010(2)
+ 2007(1)

Narrow By Add date
+ 2009(2)
+ 2010(2)
+ 2008(1)

Narrow By Channel
+ googletechtalks(2)
+ citrisuc(1)
+ fxpal(1)
+ oreillymedia(1)

Recent Searches
information
corporate capture
corporate memory
fxpal
corporate memory
[Clear]

Syndication
RSS 2.0

i4energy: Energy Research and Commercialization at SRI From YouTube
Lecture by: Barbara ...innovations. With a corporate commitment to solving important problems, this talk will cover why SRI believes ensuring affordable, clean, and dependable energy supplies is one of the (more)
Duration: 45:49 **# Slides**: 56
Channel: citrisuc **Published**: 2010/10/29 **Added**: 2010/11/01

Talkminer From YouTube
TalkMiner aggregates and indexes lecture videos available across the internet. http://www.talkminer.com
Duration: 3:50 **# Slides**: 16
Channel: FXPAL **Published**: 2010/10/21 **Added**: 2010/10/22

Situating Personal Information Management Practices within an Organization From YouTube
Google Tech Talk ...Quiñones. Personal Information Management (PIM) ...when we organize our information. This often includes ...support: encountering (more)
Duration: 56:31 **# Slides**: 49
Channel: googletechtalks **Published**: 2009/12/21 **Added**: 2009/12/29

O'Reilly Webcast: Security Monitoring From YouTube
How well does your enterprise stand up against today's sophisticated security threats? In this webcast, security experts from Cisco Systems demonstrate how to detect damaging security incidents on your global network--first by teaching you which (more)
Duration: 86:56 **# Slides**: 59
Channel: OreillyMedia **Published**: 2009/07/21 **Added**: 2009/07/21

Seamless Information Capture and Discovery for Corporate... From YouTube
Google TechTalks March ...focuses on intelligent information access, personalization and knowledge management technologies. Laurent Denoue is a senior research scientist at FX Palo Alto Laboratory. His (more)
Duration: 43:48 **# Slides**: 101
Channel: googletechtalks **Published**: 2007/10/08 **Added**: 2008/08/12

Results per page: 10

About Privacy Policy Terms of Use Contact us Copyrights FXPAL, 2008-2010

information corporate captu | Search in this talk | Search on TalkMiner | 13 matches | SHARE

Title: Seamless Information Capture and Discovery for Corporate... Original on youtube.com

Corporate Memory FXPAL
- What is it?
  - "... it allows a company to know what it knows" George Shultz, July 2004
- Without it, companies are not as efficient as they could be
  - Reuse instead of reinvention
  - Coordination instead of replication
- An important but very difficult problem

03:39 / 43:49

Link: http://talkminer.com/viewtalk.jsp?videoid=728FKMYYtJw
Channel: googletechtalks
Abstract: Google TechTalks
March 22, 2006

David Hilbert
Daniel Bilsus
Laurent Denoue

David Hilbert is a senior research scientist at FX Palo Alto Laboratory. His research interests lie in the design and evaluation of practical interactive, collaborative, and ubiquitous computing applications. Daniel Billsus is a senior research scientist at FX Palo Alto Laboratory. His

About Privacy Policy Terms of Use Contact us Copyrights FXPAL, 2008-2010

Figure 6. Left: Search results. Each lecture shows a representative keyframe, attribution, title, and description when available. Results can be filtered or sorted based on metadata including the date, duration, number of slides, and source of the webcast. Right: Viewing slides and using them to seek the embedded player.

Users enter one or more search terms and a list of talks that include those terms in the title, abstract or the presentation slides are listed as shown in the left panel of Figure 6. The information displayed for each talk in the search results includes a representative key frame, the title of the lecture, and the channel or source of the talk. Other metadata displayed includes the duration of the talk, the number of detected slides, the publication date, and the date of indexing by TalkMiner. An attribution and link to the original video source is also provided. Notice that search terms are highlighted in green to identify their occurrence.

Users can alter the sorting and filtering criteria for the resulting list of videos. By default, talks are sorted by relevance to the query terms. Other available sort attributes include publication date, number of slides, channel, and rating. The first column on the left side of the results page includes interface controls to filter results according to specific criteria (e.g., the year of publication, the channel, etc.). It also includes a list of recent search queries to allow users to re-execute a recent query. Search results link to the detailed talk view as depicted in the right panel of Figure 6. Slides matching the query are highlighted, and the user can control the playback position of the embedded video player by selecting the slide thumbnail with the content of interest. The system currently indexes over 13,000 lecture videos from a variety of sources and can be accessed at http://www.talkminer.com .

5. CONCLUSION

In this paper, we reviewed a series of research projects in multimedia browsing and retrieval conducted at FXPAL. The evolution of our research parallels trends in multimedia information creation and usage. Our earlier photo management application, the RMO, helped users move "beyond the shoebox" by facilitating content organization by automatic event clustering and interactive batch photo tagging using text or person tags. MediaGlow supports users who today require tools for photo grouping and selection for subsequent repurposing and dissemination.

In visual retrieval, our systems have steadily progressed towards increasingly heterogenous content. Our experiments examined multiple research directions including text-free, program-based, and multi-user search. Scalability has been a continuous focus with the aim of aggregating internet content; the storage footprint as well as both offline and online processing requirements are now basic design considerations.

Throughout, our systems consistently combine flexible and powerful user interfaces with established content analysis methods. We exploit automatic processing when it can reliably reduce manual interaction. When automatic methods are unreliable, we design interfaces and visualizations to enhance the user experience and maximize the utility of users' efforts. This formula underlies our systems for browsing and search of multimedia information.

REFERENCES

[1] Girgensohn, A., Adcock, J., Cooper, M., Foote, J., and Wilcox, L., "Simplifying the management of large photo collections," in [*Human-Computer Interaction INTERACT '03*], 196–203, , IOS Press (2003).

[2] Cooper, M., Foote, J., Girgensohn, A., and Wilcox, L., "Temporal event clustering for digital photo collections," *ACM Trans. Multimedia Comput. Commun. Appl.* **1**(3), 269–288 (2005).

[3] Ioffe, S., "Red eye detection with machine learning," in [*Proc. ICIP*], **2**, 871–874 (2003).

[4] Girgensohn, A., Adcock, J., and Wilcox, L., "Leveraging face recognition technology to find and organize photos," in [*MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*], 99–106, ACM, New York, NY, USA (2004).

[5] Girgensohn, A. and Chiu, P., "Stained glass photo collages," in [*Proc. UIST*], (2004).

[6] Girgensohn, A., Shipman, F., Wilcox, L., Turner, T., and Cooper, M., "Mediaglow: organizing photos in a graph-based workspace," in [*IUI '09: Proceedings of the 13th international conference on Intelligent user interfaces*], 419–424, ACM, New York, NY, USA (2009).

[7] Girgensohn, A., Shipman, F., Turner, T., and Wilcox, L., "Flexible access to photo libraries via time, place, tags, and visual features," in [*JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*], 187–196, ACM, New York, NY, USA (2010).

[8] Wilcox, L., Uchihashi, S., Girgensohn, A., Foote, J., and Boreczky, J., "Mbase: indexing, browsing, and playback of media at fxpal," in [*MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 2)*], 204, ACM, New York, NY, USA (1999).

[9] Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J., "Video manga: generating semantically meaningful video summaries," in [*MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 1)*], 383–392, ACM, New York, NY, USA (1999).

[10] Smeaton, A. F., Over, P., and Kraaij, W., "Evaluation campaigns and trecvid," in [*MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*], 321–330, ACM, New York, NY, USA (2006).

[11] Adcock, J., Cooper, M., Denoue, L., Pirsiavash, H., and Rowe, L. A., "Talkminer: a lecture webcast search engine," in [*Proceedings of the international conference on Multimedia*], *MM '10*, 241–250, ACM, New York, NY, USA (2010).

[12] Adcock, J., Girgensohn, A., Cooper, M., Liu, T., Wilcox, L., and Rieffel, E., "Fxpal experiments for trecvid 2004," in [*Proceedings of the TREC Video Retrieval Evaluation (TRECVID)*], 70–81, NIST, Washington D.C. (2004).

[13] Cooper, M., Adcock, J., Zhou, H., and Chen, R., "Fxpal at trecvid 2005," in [*Proceedings of the TREC Video Retrieval Evaluation (TRECVID)*], (2005).

[14] Cooper, M., Adcock, J., and Chen, F., "Fxpal at trecvid 2006," in [*Proceedings of the TREC Video Retrieval Evaluation (TRECVID)*], (2006).

[15] Snoek, C. G. M. and Worring, M., "Concept-based video retrieval," *Found. Trends Inf. Retr.* **2**(4), 215–322 (2008).

[16] Cooper, M. and Foote, J., "Scene boundary detection via video self-similarity analysis.," in [*IEEE Intl. Conf. on Image Processing (3)*], 378–381 (2001).

[17] Adcock, J., Cooper, M. L., and Pickens, J., "Experiments in interactive video search by addition and subtraction," in [*ACM Conf. on Image and Video Retrieval*], (2008).

[18] Baeza-Yates, R. and Pino, J. A., "A first step to formally evaluate collaborative work," in [*GROUP '97: Proc. ACM SIGGROUP Conference on Supporting Group Work*], 56–60 (1997).

[19] Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., and Back, M., "Algorithmic mediation for collaborative exploratory search," in [*ACM SIGIR*], (2008).

[20] Hauptmann, A., Lin, W.-H., Yan, R., Yang, J., and Chen, M.-Y., "Extreme video retrieval: Joint maximization of human and computer performance," in [*Proc. ACM Multimedia 2006*], 385–394 (2006).

[21] Adcock, J., Pickens, J., Cooper, M., Chen, F., and Qvarfordt, P., "Fxpal interactive search experiments for trecvid 2007," in [*Proceedings of the TREC Video Retrieval Evaluation (TRECVID)*], NIST, Washington D.C. (2007).