Semantics Meets UX: Mediating Intelligent Indexing of Consumers' Multimedia Collections for Multifaceted Visualization and Media Creation

Stacie Hibino, Alexander Loui, Mark D. Wood, Samuel Fryer, Cathleen D. Cerosaletti
Eastman Kodak Company, {hibino, alexander.loui, mdw, samuel.fryer, cathleen.cerosaletti}@kodak.com

Abstract

Unorganized media collections hinder consumers from fully experiencing and enjoying their visual media. User interfaces can mediate the results of automated indexing by presenting data and interactions that leverage the strengths of individual and combined algorithm results, supporting multifaceted browsing, and enabling user correction in a way that is not disruptive to the users' activities. We describe the Semantic System Demonstration Framework (SSDF), a flexible and extensible framework for combining multiple semantic indexing algorithms for consumer photo and video clip collections into one integrated system. We also describe key features of Koi, an SSDF desktop client application with a user interface designed to mediate and leverage the intelligent indexing incorporated in the SSDF server. Together, Koi and SSDF empower users to experience their personal multimedia in novel and sophisticated ways.

1. Introduction

Consumers want to *experience* their visual media, not *organize* them. Previous studies indicate what little organization users do with their collections is typically focused on manually grouping captured assets by date and event [1]. Algorithms for extracting semantic information from visual media can help automate organization, but such algorithms are far from perfect. Although much research has been conducted in automatically clustering assets by a particular feature (e.g., by event [2], by person [3], by CBIR [4], etc.), less work has been done in integrating several clustering algorithms into one system. While research in digital libraries has provided more integrated solutions to automatically indexing large multimedia collections, many digital library approaches leverage data typically not available in consumer-captured media (e.g., closed captioning text, video transition effects, associated titles or captions, etc.). Some novel user interfaces such as the Personal Digital Historian [5] as well as current commercial image organization software integrate browsing and search via different facets of data, but these systems heavily depend on manual input and tagging of assets by the user.

This paper focuses on addressing how imperfect algorithms may be combined in a way that maximizes the benefit to the user by using a two-fold approach to semi-automated organization of consumer photo and video clip collections:

- Build an extensible framework enabling new and updated algorithms to be easily added;
- Design user interfaces to mediate the results of automated indexing by:
  - presenting multiple views of the same data and corresponding interactions to leverage strengths of individual and combined algorithm results,
  - supporting multifaceted browsing, and
  - enabling user correction in a way that is not disruptive to the users' current activities.

Our approach is further motivated by the notion that "consumers typically put little effort into photo annotation; they are more focused on exploratory search and serendipitous discovery of photos with a stronger emphasis on entertainment" [6]. Others have also taken a semi-automated annotation approach to image organization [7], [8], [9]. Our work is differentiated from this previous work through (a) the integration of multiple views designed to leverage one or more algorithmic indexers along with corresponding facet-based browsing, (b) support for smart story creation, and (c) addition of displays for more serendipitous discovery (e.g., Drift and Connection views).

2. SSDF architecture

The Semantic System Demonstration Framework (SSDF) uses a client-server architecture. Users can upload digital images and video clips to the server using a desktop or web client. The server runs Semantic Indexers in the background to extract metadata from the uploaded assets (see Section 3). Numerous semantic indexers including event clustering, scene classification, autolabeling based on holidays, face detection and recognition, location extraction, image quality, key frame extraction, etc., have been implemented and integrated into the system.

Once users' assets have been indexed, they can use a desktop or web-based client to browse, search and/or create products with their assets. The SSDF's intelligent Story Generator automatically creates products such as calendars, photo books and digital presentations. The story generator uses semantic information and rule-based inferencing to automate the process of selecting and combining digital assets based on intended author and recipient(s).

3. Semantic indexers and algorithms

As users introduce new assets into the system, the System Manager automatically invokes the appropriate semantic indexers to run on those assets. A table in the relational database specifies which semantic indexers should be run and in what order, as some semantic indexers may rely on metadata produced by other semantic indexers. For example, the social IVI metric described in Section 3.5 relies on the people detection and recognition algorithm described next.

3.1. People detection and recognition

People detection and recognition are key algorithms for indexing consumer content. Our people clustering algorithm is the first step towards automatic recognition. Clustering images by people involves a number of intermediate steps: face detection, feature point locations, facial measurements computation, similarity computation based on facial measurements, and clustering similar faces into groups. Faces are initially detected using a fast, cascaded classifier [10]. The approximate eye locations computed by the face detector are used to initialize an active-shape model-based algorithm [3] to locate facial feature points. For each feature, the distribution of difference in value is computed for two classes—each person to same person comparisons and each person to different person comparisons—using the face recognition ground truth database. Simple Bayesian threshold-based binary classifiers are constructed for each feature based on these distributions. The AdaBoost algorithm (as in [10]) is used to determine weights to be assigned to the weak single-feature classifier outputs. For classification or clustering, a probability network is formed by establishing links between each point's *k* nearest neighbors.

3.2. Event clustering and recognition

The goal of event clustering is to automatically group images from unorganized sets of images into separate events. Within each event, images are also clustered into separate groups of relevant content called sub-events. Images in an event are associated with same setting or activity, while images in a sub-event have similar image content within an event. The event-clustering algorithm [2][11] organizes images into events and sub-events based on the date and time of image capture, and the content similarity between images. The basis of using time information for clustering is the assumption that most people arrange their photos in roughly chronological order. Moreover, time differences between images in an event (or a subevent) are typically smaller than time differences between images from different events. If date and time information is not available, then the algorithm relies on image-content information.

The goal of event recognition is to automatically classify an event cluster into one of the common consumer event categories. The classification algorithm utilizes image-level features such as people present, indoor/outdoor, and type of scene detected, e.g., nature, urban, beach, etc. In addition, event-level features are used including inter-event time, the time of day, etc. A probabilistic classifier has been constructed to automatically classify an event cluster into a number of categories such as party, sports, vacation, and family moments.

3.3. Location clustering and recognition

The goal of location-based indexing is to assign semantic location information to image and video assets. The use of location indices includes:

- Map UI for picture browsing,
- Complex text expressions (or metaphoric UI expressions) for picture searching, and
- Algorithmic support for event recognition.

There are two subareas for indexing: (1) location clustering and (2) location recognition. Location clustering automatically creates clusters of proximate images based on GPS metadata (latitude/longitude) and/or scene similarity associated with the assets. This is achieved by means of a statistical clustering of GPS metadata using the ISODATA (iterative self-organizing data) algorithm. Location recognition automatically assigns one or more referential labels to each image location or the clustered image locations. The current approach uses proximity to the nearest Geographic Names Information System (GNIS) feature as the location label.

3.4. Content-based image retrieval (CBIR)

Content-based image retrieval (CBIR) is a technique for searching for similar images based on a reference image or a region of interest (ROI) within a reference image. Color and texture are key searchable ingredients inherent with image content and logically connected to objects of interest. Three primitive feature histograms are used: color, color composition, and color texture. The current approach extracts these color and texture features from an image and indexes them in a database. During the search phase, the same types of features are extracted from the query image and compared to the indexed features. In addition to whole image search, we have also implemented a ROI search that allows the user to specify a rectangular region within the query image. ROI search will usually provide better performance over standard whole image search due to better localization of relevant object or region in the image.

3.5. Image value index (IVI)

A challenge for automated methods of content management is to derive image value predictors from image characteristics computable from the pixels, associated metadata, and user characteristics and purposes. The approach described in this work is to group these heterogeneous sources of data into a limited number of subgroups, and to determine if some combinations of measures derived from these subgroups can be used to make useful image value assessments through the use of computer learning techniques. This work primarily considers two groupings of computed image and metadata characteristics, which may form components of a limited list of measures, predicting the value of an image to the achievement of a user's purposes. The technical IVI is derived from computations of objective technical image quality parameters, and the social IVI is derived by application of a rule-based algorithm to metadata describing the social relationships between persons depicted in an image and the user of that image [12].

3.6. Scene and material classification

A suite of scene and material classification indexing algorithms will automatically detect common objects and scene types in a consumer image by its high-level meaning. These include indoor/outdoor, sky, grass, building, mountain, sunset, beach, cityscape, etc. For scene classification, the algorithms are based on a content-based image classification approach using, e.g., spatial color moments. We also consider combining content and context metadata, such as camera metadata [13]. A content-based image classification approach is used for material and object classification based on color, texture, and edge type features. We also consider combining content and spatial contextual information such as co-occurrence and spatial relationship or image regions [14].

4. Inferencing engine

The SSDF inferencing engine is a key ingredient enabling users to experience their collections in a semantically meaningful way. Inferencing is used to interpret people relationships as well as to drive the story generator. For example, if the user queries for pictures of their "mother," the inferencing engine maps the appropriate person ID to the current user's mother. Story rules are used both to determine when to produce certain products and what types of pictures should appear in a particular product.

Inferencing is carried out using Prolog-based rules executing against a third-party triplestore, AllegroGraph from Franz Inc. To enable various types of reasoning, the SSDF populates the triplestore with pertinent metadata from individual assets and events as well as metadata associated with people. The People view described in Section 5.4 provides a way for users to enter information about themselves and others portrayed in their pictures, including birthdates and anniversaries. This interface could be easily extended to include personal interests, places of residence, etc. In addition, third-party sources of knowledge may also be loaded into the triplestore, including information about holidays and GIS data. Because information about all users is stored in a common triplestore, the system can readily reason across multiple users and their collections. For example, the story generator [15] uses rule-based inferencing to determine the actual assets that should go into a particular product type.

5. Workflow and user experience

Although several SSDF web- and desktop- based clients have been developed, one desktop client, called "Koi," was specifically designed and developed to provide more comprehensive and integrated access to results of the semantic indexers, as well as to provide an interface where users could edit and add metadata to assets. Koi provides a rich user interface for multifaceted browsing and searching of a user's own media collection plus the media that has been shared via sharing groups. We highlight some of the key features of Koi in the remainder of this section.

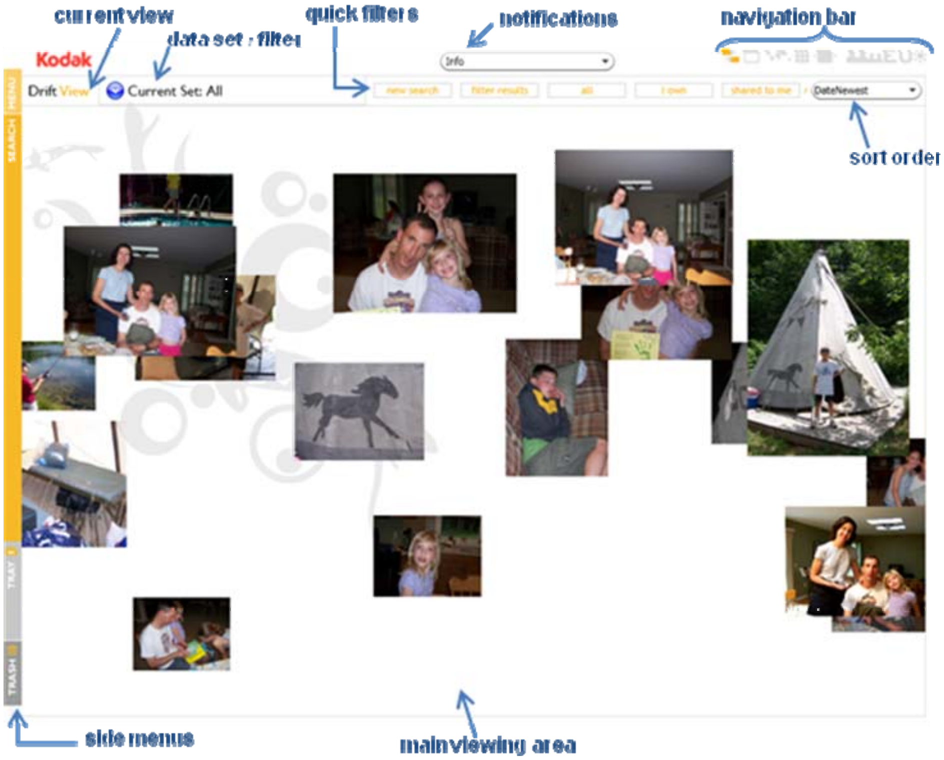


Figure 1. Annotated sample screen shot of Koi client application displaying the Drift view.

5.1. Screen layout

The main screen layout of Koi has three regions—a header area, main viewing area, and side menus (see Figure 1). Koi includes three types of views—set views, pop-up views, and other views. Users can select a set or other view via the navigation bar.

5.2. Set views

Set views provide alternative ways of viewing the same subset of media. Switching between different set views does not affect the current asset filtering. For example, if the user searches for all images containing person P, only those images will be displayed in any given set view. All set views support autoplay of video clips. When users click on an individual media thumbnail in a set view, a higher resolution version is displayed in the pop-up Single Picture overlay view (see Section 5.3). Koi includes these set views:

Drifting View: displays 12–20 media at a time, drifting horizontally across the screen (see Figure 1). The size, speed, and vertical placement of assets currently are randomly determined. Video clips autoplay without audio as they drift across the screen, adding a dimension of surprise

Slide Show View: presents a slide show of the currently filtered set of assets.

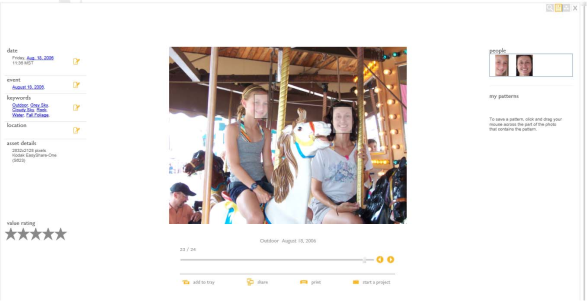
Calendar Views: collection of visual calendar-based views of different temporal granularities. Users can access: All Years (i.e., month by year thumbnail grid), Single Year, Month Over Multiple Years, Single Month, and Day calendar views. Users can drill down in time from one calendar view to another by clicking a corresponding year, month, or day.

Grid View: standard "light table" layout for efficiently viewing a large number of media at once.

5.3. Pop-up views

The Single Picture, Single Picture Details, and Connection view are "pop-up" views displayed as an overlay on top of one of the set views described in Section 5.2. The Single Picture view displays a single high resolution media item of a corresponding media thumbnail. Users can browse forward and backward through the currently selected subset of media. Video controls and key frames are provided for video clips.

The Single Picture Details view displays automatically extracted and derived metadata about the current media item (see Figure 2). The following metadata are displayed, if available: capture date, event label, keywords related to scene type, location, asset details such as image size and camera used, people (detected faces), patterns (i.e., user-specified ROIs), and image value rating. An edit icon indicates which metadata can be edited. This allows users to add details or correct errors in the metadata within the context of browsing. Users can click on metadata text hyperlinks and visual metadata such as faces. Clicking on such hyperlinks executes a corresponding query and results are presented in the Grid view. For example, if a user clicks on an event label, all media captured at that event are displayed in the Grid view.



date
Friday, Aug. 18, 2006
11:36 MST

event
August 18, 2006,

keywords
Outdoor, Grey Sky, Cloudy Sky, Rock, Water, Fall Foliage,

location

asset details
2832x2128 pixels
Kodak EasyShare-One (S623)

value rating

people

my patterns

To save a pattern, click and drag your mouse across the part of the photo that contains the pattern.

Outdoor August 18, 2006

23 / 24

Figure 2. Single Picture Details View.

The Connection view shows how the currently selected media item is related to other media based on four facets: people, event, similarity (i.e., “looks like”), and scene (see Figure 3). The system currently displays up to three related media items for each facet type. For the people facet, media are displayed that contain one or more of the people recognized in the center media item. The event facet displays media that were captured during the same event. The “looks like” facet displays media based on whole image similarity. The scene facet displays media that have the same “indoor” or “outdoor” scene value as the center media.

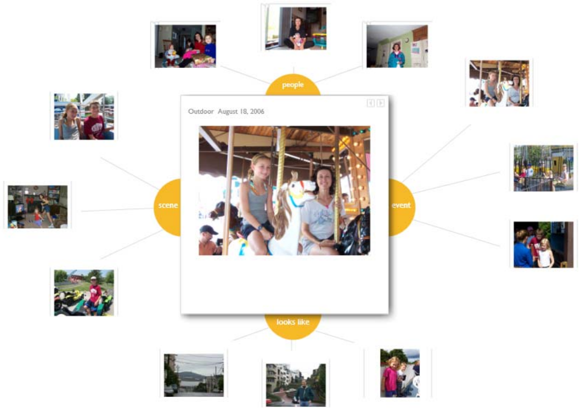


Figure 3. Connection view.

The media displayed for each facet are determined based on the results of the relevant semantic indexers and algorithms (see Section 3). When users click on any media item of any facet, that item is placed in the center, becoming the new item of focus. The corresponding facet media of the new media item are automatically recalculated and the view is updated accordingly.

5.4. Other views

The People view has three aspects: All People view, Family view, and Edit Person screen. The top half of the All People view displays one face thumbnail for each people cluster, each cluster being initially determined by the people clustering algorithm (see Section 3.1). When users click on a face, corresponding thumbnails of all assets containing the currently selected person are displayed in the bottom half of the screen. Users can merge people clusters and add or edit people’s names. Adding name labels enables searching by name and provides feedback to users when viewing pictures that contain more than one face. Users can correct any incorrectly categorized faces. The Family view shows family relationships such as spouse, parent, and children. Users can specify these relationships in the Edit Person screen. Clicking on a spouse’s, parent’s, or child’s face places that face in the center of the Family view and relationships with respect to that person are updated accordingly.

The Edit Person screen enables users to enter in profile and relationship information for individuals. Users can add information such as birthday, nickname, gender, address, email, notes, and relationship to self and others. The more information that a user provides, the more can be leveraged elsewhere by the system, (e.g., by the story generator; see Sections 4 and 5.6).

The Event view displays the current subset of assets by event, using the event clustering algorithm (see Section 3.2). Top-level events are displayed with one event per row. Each row contains event metadata, followed by sample thumbnail images from that event. The following metadata are displayed: event label, date, number of assets, and number of lower level events. An event that contains lower level events has a triangle expander button; clicking on that button expands the corresponding event, displaying all lower level events, one event per row as above.

The Everyday view for groups summarizes group-based media, activities (e.g., rating, notes), and presence. One group is placed and expanded in the center of the screen. All other groups are displayed in collapsed form around the center group. Clicking on a collapsed group swaps it into the center and expands it. In the expanded center group, the group label and currently logged in user are displayed in the center of the group. Other group members are displayed within the group circle. Shared group items are displayed along the top arc and favorite group items are displayed along the bottom arc of the group circle. Clicking on a shared event takes the user to the Grid view display of all media belonging to that event. Additional buttons provide access to all media shared to, and all favorite media of, the current group.

5.5. Menu functions

The main menu provides access to five other menus: Upload, Explore, Create, Share, and Personalize. Users can upload new media assets to the server via the Upload menu. The Explore menu provides navigational access to the views described in Sections 5.2 and 5.4. The remaining menus are placeholders for future features related to creating projects, sharing media, and personalization.

The Search menu is an interactive menu for creating a search query based on textual and/or visual metadata. Users can search their collection based on dates and events, keywords, faces, a portion of an image (e.g., to search via CBIR), and/or asset details. Users can create a logical-based AND or an OR query of the search criteria specified. The results of the search are presented in the Grid view.

5.6. Notices and media creation

SSDF alerts the user to actions it has performed on the user's behalf through the New Info drop down menu, which is accessible in all views. Currently, the system uses this mechanism to communicate to the user new products it has created on the user's behalf. For example, the story generator may observe that Mother's Day is in a week, and create a Mother's Day album for the user to view. The New Info drop down menu lists new products created by the system. The user may select a product from the list, which will result in a product preview being displayed to the user.

6. Conclusion

Since consumers may not want to manually organize their visual media, automatic mechanisms for visualizing media are required to provide a satisfying user experience, particularly as the size of media collections increases to include thousands of assets. As imaging algorithms remain imperfect, the user interface must mediate the results of automated indexing. This paper describes a new approach to integrating semantic indexers and algorithms into a flexible, semantically aware client-server architecture. The SSDF architecture and Koi client allow for multifaceted search, browse, and creation based on the metadata generated from the output of a variety of indexers and algorithms. Koi employs a novel user interface for searching and browsing collections using a variety of visualization types. These different views each draw on different semantic information, giving the user flexibility in selecting the most appropriate view, depending on both the user's intent and the availability of semantic information. Potential future work includes evolving existing algorithms to increased performance levels as well as integrating new algorithms developed in concert with user testing.

Acknowledgements. Many people contributed to the development of the SSDF. We especially thank Pat Cosgrove, Dale McIntyre, and the Extractable team.

7. References

[1] D. Kirk, A. Sellen, C. Rother, and K. Wood. "Understanding Photowork." In *CHI 2006 Conf. Proc.* ACM, New York, NY, 2006.

[2] Loui and A. Savakis, "Automated Event Clustering and Quality Screening of Consumer Pictures for Digital Albuming," *IEEE T. Multimedia*, Sept 2003.

[3] M. Bolin and S. Chen, "An Automatic Facial Feature Finding System for Portrait Images," *Proc. of IS&T PICS*, 2002.

[4] T. Deselaers, D. Keysers, and H. Ney. "Features for Image Retrieval: An Experimental Comparison," *Information Retrieval*, 11(2), Springer, 2008.

[5] C. Shen, N. Lesh, and F. Vernier. "Personal Digital Historian: Story Sharing around the Table." *interactions*, 10(2), Mar. 2003.

[6] B. Shneiderman, B. B. Bederson, and S. M. Drucker. "Find that Photo!: Interface Strategies to Annotate, Browse, and Share," *Commun. ACM*, 49(4), April 2006.

[7] Kuchinsky et al. "FotoFile: A Consumer Multimedia Organization and Retrieval System." In *CHI '99 Conf. Proc.* ACM, New York, NY, 1999.

[8] Suh and B. Bederson. "Semi-Automatic Image Annotation Using Event and Torso Identification," *HCIL-2004-15*, Univ. of Maryland, College Park, MD, 2004.

[9] M. Tuffield et al. "Image Annotation with Photocopain." In *SWAMM 2006 Workshop Proc.*, 2006.

[10] P. Viola and M. Jones, "Robust Real-time Face Detection," *Int. Journal of Computer Vision*, 57(2), 2004.

[11] C. Cerosaletti, M. Das, A. Loui, and B. Kraus. "Approaches to Consumer Image Organization Based on Semantic Categories," Proc. SPIE's *Intl. Symp. on Optics East–Multimedia Systems and Applications IX*, Boston, MA, Oct. 2006.

[12] Loui, M. D. Wood, A. Scalise, and J. Birkelund. "Multidimensional Image Value Assessment and Rating for Automated Albuming and Retrieval," *IEEE Intl. Conf. on Image Processing* special session, Oct. 2008.

[13] M. Boutell, J. Luo, X. Shen, and C. Brown. "Learning Multi-label Scene Classification," *Patt. Recog.*, 37, 2004.

[14] Singhal, J. Luo, and W. Zhu. "Probabilistic Spatial Context Models for Scene Content Understanding," Proc. *IEEE Computer Vision and Pattern Recognition*, 2003.

[15] M. D. Wood. "Exploiting Semantics for Personalized Story Creation," *Proc. 2008 IEEE Intl. Conf. on Semantic Computing*, Aug. 2008.