Reliving On Demand: A Total Viewer Experience

Vivek K. Singh[1*], Jiebo Luo[2], Dhiraj Joshi[2], Phoury Lei[2], Madirakshi Das[2], Peter Stubler[2]
[1] University of California, Irvine, [2] Kodak Research Laboratories, Rochester, NY, USA
singhv@uci.edu, {jiebo.luo, dhiraj.joshi, phoury.lei, madirakshi.das, peter.stubler}@kodak.com

ABSTRACT
Billions of people worldwide use images and videos to capture various events in their lives. The primary purpose of the proposed media sharing application is digital *re-living* of those events by the photographers and their families and friends. The most popular tools for achieving this today are still static slide-shows (SSS) which primarily focus on visual effects rather than understanding the semantics of the media assets being used, or allowing different viewers (e.g. friends, family, who have different relationships, interests, time availabilities, and familiarities) any control over the flow of the show. We present a novel system that generates an aesthetically appealing and semantically drivable audio-visual media show based on several reliving dimensions of events, people, locations, and time. We allow each viewer to interact with the default presentation to 'on-the-fly' redirect the flow of reliving as desired from their individual perspectives. Moreover, each reliving session is logged and can be shared with other people over a wide array of platforms and devices, allowing sharing experience to go beyond the sharing of the media assets themselves. From a detailed analysis of the logged sessions across different user categories, we have obtained many interesting findings on the reliving needs, behaviors and patterns, which in turn validate our design motivations and principles.

Categories and Subject Descriptors
H.5.2 [INFORMATION INTERFACES AND PRESENTATION]: User Interfaces - Interaction styles. H.5.4 [INFORMATION INTERFACES AND PRESENTATION]: Hypertext/Hypermedia-Navigation

General Terms
Design, Human Factors.

Keywords
Reliving, sharing, semantic, interaction, slide-show, spatio-temporal navigation

1. INTRODUCTION
Human beings have always felt the need to chronicle parts of their daily lives. From the paintings of cave men, to oral ballads and story-telling, to commissioned portraits in the medieval era, we notice a clear human desire to record and preserve aspects of their lives to relive at later points in time. The 20th century saw the emergence of personal cameras to undertake this chronicling process, and today nearly every family chronicles parts of their lives using photos and videos. This chronicling and data capturing was done for two important reasons. First, for the participants themselves to *relive* the events of their lives at later points of time; second, for them to *share* these events in their lives with other friends and family, who were not present at the events but would still be interested in knowing how the event (e.g. the vacation, or the wedding, or the trip) went by. The importance of sharing is underscored by the *billions* of image uploads per month on social media sites such as Facebook, Flickr, and Picasa. As shown in Fig. 1, digital reliving has become ubiquitous on many platforms and devices.

Figure 1. People don't merely share images, they want to re-live and share their experiences with others. It is important to enable digital reliving on ubiquitous platforms (web, home, hardcopy) and devices (TV, PC, smart phones, tablets, digital picture frames, kiosks, photobooks).

The massive growth in data *creation* aspect has highlighted two major issues with the techniques available for *consumption* of such data. First, the sheer volume of images implies that it becomes difficult for participants to relive specific memories (e.g. in terms of events, locations, people), without searching through a huge collection of media. Second, sharing one's images with multiple family members and friends in one's diverse social network implies that each person has a different interest level and perspective when looking at the images. Sharing across different users requires catering to users with different time availability, different motivations, different interests, and different perspectives. These different perspectives have traditionally been ignored by both research and commercial applications. In fact, while sophisticated tools are being created to bridge the *semantic gap* on the organizer or the producer side of media collections, very little effort is being put into bridging the *intent gap* on the *recipient* side of the media sharing equation.

* Work was done when the author was interning at Kodak Research Laboratories, Eastman Kodak Company, Rochester, NY, USA.

*Area Chair: Dick Bulterman

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

Considering these factors, we present a novel system for digital *reliving* and *sharing*. We describe an approach for generating *dynamic* media-shows (which are inspired by, but significantly richer than *static* slide-shows which are currently the most popular mode of photo playback). These media reliving experiences are aesthetically pleasing yet semantically drivable based on people, locations, events, and time - attributes that have been successfully employed by multimedia community to characterize stories in social media settings such as Facebook [16]. Very importantly, we no longer consider the recipient to be a passive consumer, but rather someone who can interact with and redirect the flow of this media-stream based on different semantic facets. This redirection occurs in an aesthetically pleasing manner and on-the-fly (i.e. without the need to re-start the show or re-compiling the data collection). Further, the users can directly interact with the show (e.g. by clicking directly on 'Hotspots' i.e. any of the faces being shown) rather than having to use a search box with "advanced" settings which appear inelegant and may disrupt the media flow. Thus, the proposed system provides users an elegant tool for reliving their personal, or their friend's media collection while interacting directly with it as and when they desire, to reroute the media flow along any desired semantic axis. As will become clearer, media reliving is different from both media search and media browsing. Media search involves clear user intent. The intent is significantly weaker in media browsing [6]. Media reliving provides a valuable middle ground for user interaction and intent clarity, and does so in an aesthetically pleasing manner.

The design of the system and its evaluation by users with different roles and demographic backgrounds (e.g. age, gender, family roles) also provided some unique insights into the process of digital re-living. While most viewers appreciated the ability to semantically re-route the media stream, their methods of interacting with it varied significantly across different demographic factors. Hence we also present our findings about the semantic axis (time, location, person, and events) that is most frequently used for rerouting the media stream, and how this distribution changes across different demographic groups. The findings are interesting, and may affect the design of ours as well as other reliving efforts in near future.

The outline for the rest of this paper is as follows. Section 2 describes the related work. Section 3 describes our design principles. We discuss the proposed reliving approach in section 4. Section 5 discusses implementation details, while section 6 describes our in-depth analysis and findings on the user feedback, as well as user interaction patterns. We present out conclusions in Section 7.

2. RELATED WORK

A number of commercial products provide similar features. For example, iPhoto, Picasa, Flickr, and Facebook all provide ways to upload and share pictures. However, the slide-shows supported are basic, with little to no user interaction present at *run time*. Hence they do not allow a user to bridge the intent gap. The FaceMovie feature in Picasa is an interesting innovation that tries to assemble images of the same person at different ages in a dynamic slide-show. Therefore, in a way it tries to understand the face/person semantics but again does not provide any user interaction tools at run time.

The multimedia community has seen multiple efforts in understanding the content of the images and bridging the semantic gap. For example, very good surveys of work in this direction are provided in [19][6]. Earliest attempts to understand the user intentions at run time were documented by "relevance feedback" work [18] that allows users to click on a desired image to see similar images. Recently multiple approaches like 'exploratory search' [15], 'faceted browsing' [24] have also started guiding the user towards their desired content. We draw inspiration from these works, but focus on supporting direct user interaction with an already running aesthetically pleasing reliving experience.

The most common way of reliving image collections till date has been static slide-shows. Notable research efforts have been made to change this paradigm [3][5][23]. Tiling slide-shows [4] create dynamic (multiple tiles of images at the same time) slide-shows with (matching beat) music to improve user experience. iSlideshow [5] aims at understanding the 'content' of the images and creates collages of images based on concepts and supports transitions that dissolve at face ROI (region of interest). Again while these works show a clear research trend in understanding the semantics of media for better slide-shows, they do not provide any user-interaction or handle the intent gap. In [23], songs are identified to accompany image collections in media shows.

Hardcopies are also useful for reliving. HP's photo collage effort [22] allows for creation of aesthetically pleasing collages of pictures. However it focuses on static collages rather than dynamic audio-visual media experiences. In another recent work that focuses on photobook creation from social media [16], media in one's social network is matched with text queries (where, when, what, who) and then arranged into a printed photo book. Works like [1] provide a rich media authoring tool set to define media shows. However, the show needs to be defined *before* (i.e at 'compile-time') the user can view it. Similarly, the system in [3] allows viewers to edit videos for shared (or personal) consumption later. In contrast, we focus on providing an effective tool that allows users to redirect the flow *while* they are viewing it (i.e. at run-time).

User interaction is a key aspect of human centered multimedia in general and is particularly important to media sharing and reliving. The recent emphasis on interactive search and browsing (closely related to our work) bears testimony to the importance of human in the loop. Along these lines, the interactive search engine in MediaMill [20] allows users to rapidly review video content and revise their search strategies on the fly. For mobile video browsing, in [8] a three-level design space consisting of video segments, entire video, and collection of related videos is proposed to characterize the complexity of navigation.

The multimedia community has recognized that human centered computing systems should be "multimodal, proactive, and easily accessible to a wide range of users" [9]. We take to heart this philosophy in defining the design principles of our work.

3. DESIGN PRINCIPLES

We postulate that reliving systems should be *user controllable*, *semantically drivable,* and *aesthetically pleasing.* These principles have had multiple design implications for the system developed.

1) User controllable: To bridge the user intent gap we allow viewers to interact with, and redirect the media flow as and when desired (users can of course also just choose to sit back and enjoy the ongoing media flow).We support *interaction on the fly,* allowing users to interact with the system and redirect the flow without causing it to stop and re-boot. This is significant as we believe that systems that stop and reconfigure each time a new input setting is desired cannot provide satisfying reliving experience.

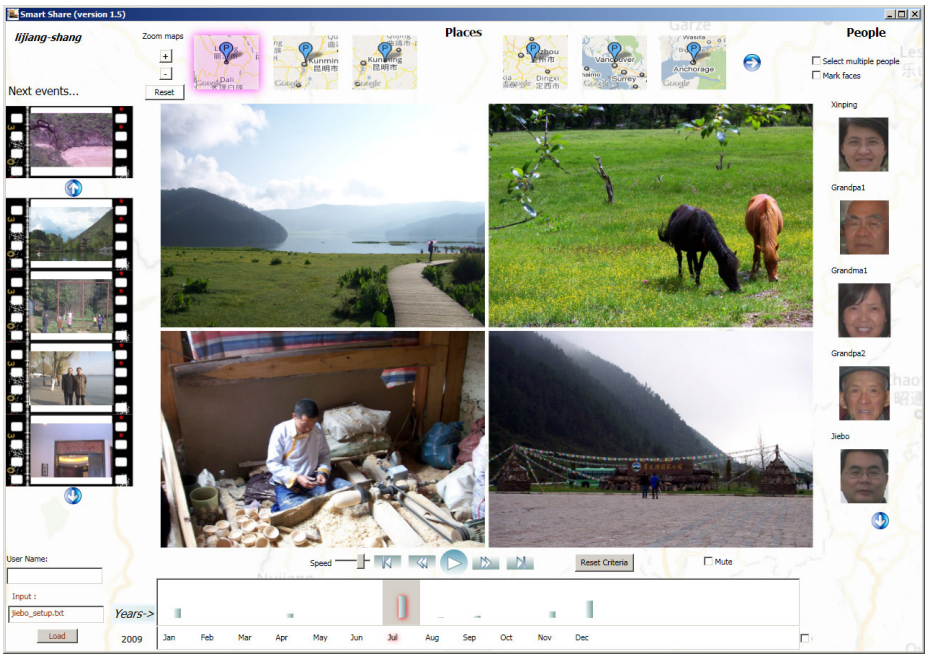

Figure 3: A blow-up of the media reliving interface showing all the features.

2) Semantically drivable: The user control cannot be just linear (e.g. pause, play) but rather needs to be semantically drivable. We consider *media (image, video) as portals to memory and experience space.* Users are interested in reliving significant events in their lives rather than viewing "IMG_0667.jpg" per se. Hence we want to provide users the tools to navigate their experience space (and not merely the media space). Specifically, we leverage on *events as organizing units* (as the human memory is believed to be largely episodic i.e. event driven) [10]. To support the experiential navigation, we follow the adage of *"less is more"* and focus only on well understood semantic axes of *time, location,* and *people* which can also be detected quite robustly. We make a design choice to keep a minimalistic number of axes to keep browsing easy yet robust.

3) Aesthetically pleasing: Aesthetics of the presentation are just as important as the media assets selected in the reliving experience. Hence features such as reflective music, background and transition effects are integral parts of the system designed. We support *dynamic presentation* i.e. present fast-paced multi-photo composites which enhance holistic event level reliving, and produce a more lively experience. Lastly, rather than focusing on mono-media (e.g. just images) we *support multimedia.* We think that the users employ more than one media (images, videos, music) to capture their lives, and enjoy a better reliving experience when multiple media are presented to them synergistically.

4. OUR APPROACH

4.1 System overview

Fig. 2 provides an overview of our media reliving system. The system pre-processes the media collection for each user to extract different event and geographical clusters, and obtains their corresponding metadata (location, time, and people). The combination of media collection and the metadata is next used to create an aesthetically pleasing media show. At run-time however, the viewer can choose to interact with the system, and the presentation is dynamically adapted to match user intent. Specifically the presentation order of events and the relative screen presence given to each photo within the event are re-computed based on user input. We also maintain a log of user-interactions, which allows us to identify different patterns as discussed in section 6.2.

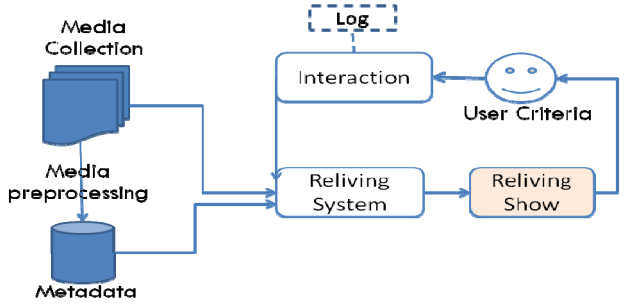

Figure 2. The high-level design of our reliving system.

A screenshot of the developed system is shown in Fig 3. As can be seen, the user has the option to passively view the default media show, or actively control the flow through the semantic axes of places, people, and time. At any given event the user is also provided the option to preview the next few events and jump forward or backwards. The user can also control the pace of the slide-show and pause to look at important images if desired. Lastly, the system can automatically go into 'full-screen' mode and back to show and hide the controlling toolbars depending on whether the user is interacting with the system.

4.2 Media reliving flow

The steps required for dynamic re-ordering and presentation reconfiguration are shown in Fig. 4. The system performs the pre-processing to cluster the media collection into events and extracts the corresponding metadata. The system chooses 'time' as the default criterion to generate the media show. However, each time

a user selects a new criterion, the system reorders the event collection based on the criterion. Next, it determines the suitability of each image inside the event media sub-collection, based on the criterion. Depending on the number of relevant images found suitable, a presentation layout is selected. Similarly, based on the criterion, the transition method across images and background music are selected. The image set is shown in order and the show continues until a new criterion is selected by the user (in our current version, the show loops back to the beginning when all the pictures corresponding to a user-criterion have been displayed). In the following sections, we describe each component in more detail.

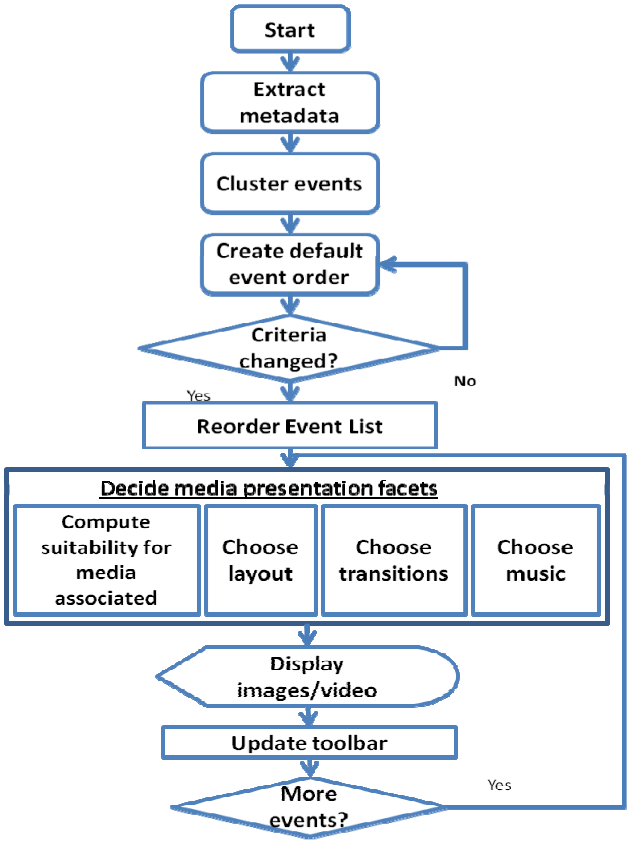

Figure 4. A flow diagram of the reliving system.

4.3 Extracting metadata (Pre-processing)

As shown in Fig. 5, the system computes three types of metadata for each image/video in the media collection.

1) Media descriptors: i.e. Type (photo/video), URL, and resolution.

2) Aesthetic value descriptor: This was obtained using an image value descriptor based on some of the attributes described in [12][11][7] including colorfulness, contrast, sharpness, spatial distributions of edges and colors, and faces.

3) Semantic metadata: (Timestamp, location, and people): The location was obtained from the combination of directly encoded GPS values, and folder directory names at coarse level. The folder names were geo-coded using Google API. The people pictured in a collection were detected using the Omron face detection algorithm similar to [21]. We developed an algorithm that uses facial similarity to automatically group faces into clusters. Face clusters were labeled once for each data set.

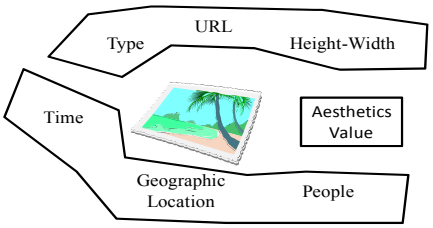

Figure 5. Metadata computed for each media element.

4.4 Clustering events

Once the system has the time-stamps for each media element, it performs event clustering using the algorithm in [13]. This algorithm is based on temporal information and visual similarity, attempting to match user's perceptions of real-life events. The histogram of time differences between adjacent images or videos is clustered into two classes: time differences that correspond to event boundaries, and those that do not. Color block-based visual similarity is used to refine the event boundaries. Once the media elements are clustered into events, the system computes the aggregated time-span, location-span and people list for the event.

The overall process of meta-data extraction for identifying the current toolbar options to be provided is shown in Fig. 6. From the geographical location axis perspective, the system clusters the data into geographical locations from which events are captured. The geographic locations computed are used as options on the 'places' browsing axis. The geographical clustering was performed using a mean-shift algorithm similar to [2].

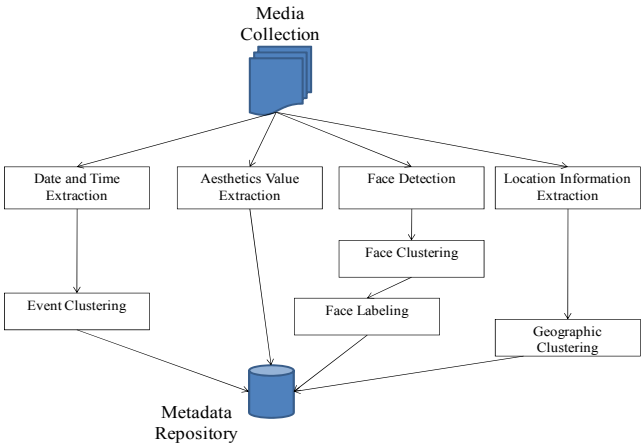

Figure 6. The metadata extraction process.

4.5 Reordering of the event list

The system uses time as the default ordering criterion for an event collection. Hence the show moves in a chronological order until the user chooses to interact with it.

The event list can be reordered according to different criteria. Each event is granted a matching score (α_1) based on the criterion selected. This matching score is combined with the previous matching score (α_2) using a weighted averaging scheme.

$$\alpha_{1(t+1)} = w_1 \cdot \alpha_{1(t)} + w_2 \cdot \alpha_{2(t)}$$

While the associated weights (w_1, w_2) ratio was heavily biased towards the new match, the presence of a dampening factor helps to create a smoother transition across criteria, and also adds some randomness (i.e. the users do not see *exactly* the same show each

time a particular value is chosen) to the created show. The precise computation is as follows:

1) Time: If the criterion selected is a particular time value (v_1), the matching criterion w_1 is defined as the normalized time difference between v_1 and the considered event's start time. We decided to go only in one direction (i.e. chronologically increasing) as an initial feedback from users suggested that they become confused by jumping back and forth in time. Hence for an event **e**, the matching score is computed as:

$$w_1 = 1 - \frac{e.startTime - v_1}{\mathrm{e_{last}}.\mathrm{startTime} - v_1}$$

where $\mathbf{e}_{\text{last}}$ is the event with the latest start time. All events with start times before the selected value are assigned a w_1 value of zero.

2) People: If the criterion selected is a person (or a group of people), w_1 is computed based on the number of that person(s) images (both in ratio and absolute numbers) in the event media collection. Therefore, if a user clicks on "Jenny's" face, the event collection that contains a large number of images, and most of which have Jenny's face is likely to attain a high matching score. Thus for selection of users p_1 *through* p_n, the matching score is computed as:

$$w_1 = 0.5 * \frac{e.numPics(p_1 \square \ldots \square p_n)}{|\mathrm{e.numPics}|} + 0.5 * \frac{|e.numPics|}{|\mathrm{e_{max}}.\mathrm{numPics}|}$$

where $\mathrm{e_{max}}$ refers to the event with the largest photo collection.

3) Location: If the user selects a particular location (v) as their criterion, the matching score w_1 is computed as the normalized distance between the chosen location and the event's geo-location. We have used [latitude, longitude] pairs to represent the centroid location of events.

$$w_1 = \frac{\|v - e.loc\|}{\sqrt{180^2 - 90^2}}$$

where the normalization has been undertaken based on the maximum possible distance in the lat-long representation space.

4.6 Deciding media presentation facets

Once the system has an ordered list of events whose images and video need to be presented in the reliving session, it proceeds to compute a score for each individual image/video. This score dictates which images are shown and determines their screen presence (time and screen real-estate).

4.6.1 Computing score for each image

The relevance of each image in reliving is computed as a weighted average of its aesthetic properties as well as semantic aspects (where relevant). We next describe score computation with respect to different reliving dimensions.

1) Time & Location: Images receive a score purely based on their aesthetic value metadata computed during the pre-processing step.

2) People: The images are given a score based on the combination of multiple semantic factors based on the people selected (i.e. p_1 through p_n) in the criterion

a). r1=$(p_1 \square \ldots \square p_n)$ present in the image? (A binary 1, 0 score)

b). r2= size of the face(s) in pixels

c). r3= position of the face centroid

d). r4= $\frac{|(p_1 \square \ldots \square p_n)|}{|total\ NumFaces|}$

e). r5= aesthetic value factor

(a) two images (b) three images

(a) four images (b) five images

Figure 7. Examples of different layout templates.

Figure 8. 'Default' layout template (left) and a template with mixed landscape and portrait frames (right).

4.6.2 Choosing the layout

Ideally the layout should be aesthetically pleasing but relevant to show the media (images or videos) in a given event. For a given event, the layout is selected such that the number of images or video shown is an integer factor of the total number of images or videos in the event. The page layouts with two, three, four or five images are pre-designed, as shown in Fig. 7. If the number of images is not a direct factor, a default layout as shown in Fig. 8 is selected. A light weight automatic cropping algorithm (see Section 4.7) is used if necessary to fit the images into a template. A set of rules is also used to determine whether to use a template with more landscape frames or portrait frames.

4.6.3 Choosing the transition

We wanted to create a dynamic experience for users viewing the images. Hence instead of changing single images (or even collections) one by one, we decided to allow each image/ frame inside the 'image collage' to transit dynamically. The actual transition method depends on the reliving criterion. It was selected to be slide-in/slide-out for the 'time' and 'location' criteria. For the people based interaction, we chose to implement a novel face-to-face transition effect. Our approach gives an effect of the person's face being used as a 'wormhole' to progress from one image or video to another. Note that our transition effect is different from Picasa FaceMovie wherein a subsequent picture bearing a face of the same person is superimposed on top of the current picture to perfectly align the face in size and position.

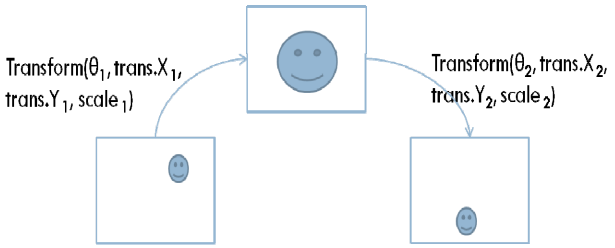

Figure 9. Creating face-to-face transitions.

As shown in Fig. 9, an affine transformation (rotation, translation, and scale) is computed to move the current face in the current image or video into a predefined size-orientation intermediate frame. At the same time a transformation to move from the intermediate frame according to the face into the next image is also computed. When both these transformations are applied in tandem as a smooth transition, the above mentioned face-to-face transition effect is produced.

We consider this an interesting feature as most currently available presentation mechanisms do not create any transitions based on the semantics of the images being shown.

4.6.4 Choosing the music

This step selects semantically relevant music to accompany the media to be displayed. The music is selected based on the current user criterion chosen. For the 'time' criterion, music is selected based on the season (e.g. music for seasons spring, summer, autumn, and winter) of the event; for criterion 'people', music is selected based on the generation of the person(s) selected (e.g. music for the generation of Baby-boomers and generations X, Y, and Z); for criterion 'location' the system searches into a database of geolocation-characterized music and chooses the one that is closest to the location of the event. A database of geo-characterized music has been constructed by manually searching for music using location specific text queries in an annotated music database (here YouTube). Locations for which characteristic music was obtained were identified based on popular tourist destination locations. A library of close to 100 songs exists in the current system. The library is extensible but that is beyond the scope of this paper.

Note though that during reliving, an event can be arrived at using one (or a combination of multiple) user criteria. In other words, an event has multiple facets (e.g. people, time) associated with it. The accompanying music is based on the current facet chosen to guide the media show. Hence the media for the same event may be accompanied by different types of music depending on how the event is arrived at in different or even the same reliving sessions.

4.7 Presenting the media

Once computing the suitability, layout, transition, and choosing the accompanying music for media is accomplished, this step displays the images or video, one event at a time. The images or video are granted screen time based on their suitability score (as described in Section 4.5). Within an event, images and video are displayed in a purely temporal order so as to present a clear temporal flow to the reliving user. The images that did not meet the minimum threshold (i.e. were semantically or aesthetically not satisfactory), are not included in the reliving session.

At times, the orientation of the images or video (landscape or portrait) may not match with the screen space allotted to them based on the selected layout. In order to resolve this issue, the system performs a light weight auto-zoom-crop to maintain as much semantic content as possible. For the criterion 'people', the auto-zoom-crop attempts to preserve the chosen person's face in the center of the frame. If the criterion is not 'people' and images or video contains faces, the auto-zoom-crop attempts to preserve as many faces as possible. If no faces are found in images or video, auto-zoom-crop attempts to conserve the center portion of images while sacrificing the outer parts.

In order to avoid the problem of having multiple images or video transitioning out of the screen at the same time (and leaving large blank screen spaces), we have implemented a token passing scheme between different holes, which allows only one image or video frame/hole to be blank on the screen at a given time.

The system also creates or selects a background image based on user selection (for time criterion, a representative image is selected from the current event, for people criterion, a mosaic of faces in the collection is created as background, and for location criterion, a location snapshot from Google maps is used as background). The background shows-up at the start of the event and compensates for empty holes during transitions. Background images also provide the users a sense of transition between events as they change during reliving.

4.7.1 Miscellaneous navigation and playback controls

As shown in Fig. 3, we also provide a number of navigation controls. The arrow at the end of the place or people navigation bar can be used to bring in new icons.

In particular, for the events, a "preview" function is provided for a user to scroll through ranked events without interrupting the current media flow. This is done by showing a sample image or video of the event as a representative thumbnail and giving (mouse-over) details of the event type and number of images or videos in them. This is akin to "channel surfing using the picture-in-picture feature in TV.

In addition, a set of intuitive playback controls (play/pause, ffwd, fbwd, next event, previous event, playback speed) are provided. Occasionally, the users want to merely control the temporal flow of their reliving experience. Hence we allow them to increase/decrease the speed of playback and pause at important events as they desire. Similarly, sometimes the users may want to 'skip' or 'rewind' events in the show.

4.7.2 Video integration

Videos are played along with the associated photos. Note that we manage the audio playback such that the ambient music is on mute when the video has its own sound track.

4.8 Updating the toolbars

The user navigation toolbars are updated based on the content of the event being displayed. The 'time' browsing toolbar is simply highlighted with the month or year of the current event, to establish a timeframe. Other aspects of this toolbar remain constant and allow users to switch to any time instance at will. The 'people' toolbar shows people relevant to the event being displayed. The 'location' toolbar shows locations of nearby events. The 'people' and 'location' toolbars also contain certain slots that, if requested, could be populated by "people" or "locations" randomly chosen from the user collection. The rationale behind this randomization is to allow the user to take a random leap in the reliving experience space if they get bored with the current selection.

4.9 Logging viewer sessions

We record various details of user reliving sessions. This serves two purposes. Firstly, it helps us in our analysis of user behavior across different demographics. Secondly, we expect users to consider sharing their reliving sessions with others as a form of experiential media. The idea is akin to sharing a book with others which already has underlines, highlights, or a Facebook/Youtube media element which shares the 'likes' or comments on it, or sharing one's play-list in addition to the actual song collection.

The recording for this purpose is done in an XML formatted log file, which captures the details of the event, the associated media elements being shown, the layout, as well as the contents of the axes being displayed. In addition, the type of click (hot-spot or axis) and the time-stamp are recorded. Thus such a record contains enough details to re-play or re-enact an entire reliving session if required. Fig. 10 shows a segment of a sample log file.

```
<Interaction>
        <Click>
                <GlobalEventID>urn:guid:f1337996-3c28-4345-b4fb-c4f1b788fc05</GlobalEventID>
                <SortedEventID>0</SortedEventID
                <TimeStamp>10:17:47 AM</TimeStamp>
                <Criteria_type>gps</Criteria_type>
                <Criteria_value>61.2175937710438 , -149.898739309764</Criteria_value>
                <HotSpotClick>False</HotSpotClick>
        </Click>
        <Snapshot>
                <Locations>
                        <loc>-149.898739309764,61.2175937710438</loc>
                        <loc>-73.508556462585,40.5956603174603</loc>
                        <loc>102.757525301205,25.1018832329317</loc>
                        <loc>104.195397,35.86166</loc>
                        <loc>6.09306585111111,52.7236709366667</loc>
                </Locations>
                <People>
                        <peo>Jiebo</peo>
                        <peo>Joyce</peo>
                        <peo>Xinping</peo>
                        <peo></peo>
                        <peo></peo>
                </People>
                <SortedEvents>
                        <eve>urn:guid:f1337996-3c28-4345-b4fb-c4f1b788fc05</eve>
                        <eve>urn:guid:f1337996-3c28-4345-b4fb-c4f1b788fc05</eve>
                        <eve>urn:guid:f1337996-3c28-4345-b4fb-c4f1b788fc05</eve>
                        <eve>urn:guid:f1337996-3c28-4345-b4fb-c4f1b788fc05</eve>
                        <eve>urn:guid:f1337996-3c28-4345-b4fb-c4f1b788fc05</eve>
                        <eve></eve>
                </SortedEvents>
                <PicsShown>
                        <pic>c:\data\jiebo\cvpr2008\103_5972.jpg</pic>
                        <pic>c:\data\jiebo\cvpr2008\103_5973.jpg</pic>
                        <pic>c:\data\jiebo\lijiang-shangrila-day2\108_0043.jpg</pic>
                        <pic>c:\data\jiebo\lijiang-shangrila-day2\108_0044.jpg</pic>
                </PicsShown>
        </Snapshot>
</Interaction>
```

Figure 10. A partial snapshot of a sample XML to record different interactions.

5. IMPLEMENTATION DETAILS

The system for reliving media collections has been implemented in C# language using Visual Studio 2010. Multiple features from the Windows Presentation Foundations were employed to make implementation easier when dealing with multiple media (image/video) types and undertaking animation effects on them.

The implemented system employs multiple multimedia processing techniques (viz. face detection, face recognition, event clustering, image geo-clustering, aesthetics value detection) under the hood. We have tried to adopt and customize many existing algorithms by computer vision and multimedia communities wherever appropriate rather than 're-inventing the wheel' ourselves.

The images used for the experiments were downloaded from web albums (Picasa, Flickr) of our volunteer participants. The pre-processing (face clustering, event detection, location clustering) was automatically undertaken by the system. The participant input was required at two time instances: once in labeling the clustered people faces, and second (if necessary for non-geotagged data) to label the media folders with a geo-encodable name.

6. EXPERIMENTS

We conducted two sets of experiments to study the performance of our created system. In the first experiment we compared the user satisfaction with our system to their favorite reliving software (e.g. iPhoto, Picasa, Facebook, ACDSee). We also studied the effect of demographic factors on the choice of features by the users. We tried to study which axes and methods of interaction were adopted by the users and how they varied over different gender and family roles.

Both the experiments were conducted using 11 family media collections. Some statistics are shown in the table below. Each collection was pre-processed and made available for users including 1st (owner of collection) 2nd (immediate family member of the collection owner, e.g. spouse, parents, or children) and 3rd parties (friends, other relatives, and acquaintances of the collection owner) to undertake their reliving sessions.

Age of contributing photographers	23 to 56
No. of images/ videos in the collection	2,091 to 10,522
No. of calendar years in time span	3 to 10
No. of tagged people in the collection	26 to 137
No. of places in the collection	19 to 45

We recorded a total of 35 reliving sessions undertaken by 26 different (14 male, 12 female) participants who interacted with the system in first (11 times), second (13), and third party (11) roles. Some participants undertook multiple roles (e.g. once as 1st party and once as a 3rd party for some other collection). However there was no overlap between 1st and 2nd party roles, as the participants belonged to 11 different families and their friends.
Each of the participants was provided with a laptop containing their pre-processed media as well as the reliving system. Each participant was requested to spend at least 15 minutes in a reliving session in their home or office settings. The average time recorded for each session however was 30.14 minutes.

6.1 Experiment 1: Comparison with commercially available options

In this experiment we interviewed 20 of the aforementioned participants and asked them for feedback on the system used. The aim of this experiment was to study the usefulness of our new system (called 'Relive!') as compared to their current favorite system. The current favorites of users included Picasa, iPhoto, and many others (e.g. Facebook, ACDSee). The users were asked to rate the systems based on 7 criteria listed in the following table. The first three criteria jointly correspond to the design principles '*user controllable'* and *'semantically drivable'* described in section 3. Next two questions correspond to the *'aesthetically pleasing'* design principle. Lastly we asked the users to rate their overall quality of experience and mention their feature wish-list.

EVALUATION CRITERIA
1. Experience Control: Did you feel that you could redirect the flow of the experience as / when desired (on the fly)?
2. Personalization: Did you feel that you could obtain a personalized reliving experience from this system?
3. Reminiscence: Did you feel like you were reminded of the relevant events in your life?
4. Liveliness: How exciting and lively did you find the experience?
5. Aesthetics and Ambience: Was it a pleasant and rich experience?
6. Overall Quality of Experience: How happy were you with the experience?
7. Feature wish-list: What new features do you most wish to add?

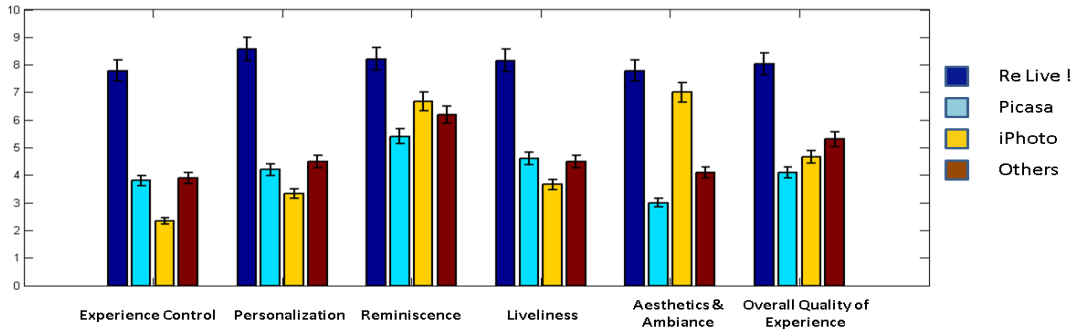

Figure 11. Comparing user-ratings for Relive! with existing multimedia organizer and viewer systems. The colored bars represent the average values and the error bars at the top indicate the 90% confidence intervals.

A summary of average user ratings (between 1 and 10) for the 6 quantitative questions is shown in Fig. 11. It is clear that the users rated the proposed media reliving system significantly higher in all six criteria . The Relive! system outperformed others most significantly in terms of '*experience control*' and '*personalization'* which resonate strongly with our design principles of creating a *semantically drivable* and *user-controlled* reliving system. Amongst other systems used, the users rated the aesthetics and ambiance of reliving sessions in iPhoto the highest.

6.2 Experiment 2: Use of different features across different user demographics

We conducted experiments to study the relative importance of different interaction axes, and how this varied across different demographics. Gender and Participation Role were chosen to be the two fundamental axes for analysis. Participation Roles were defined as aforementioned 1st party, 2nd party, and 3rd party.

As previously mentioned we recorded a total of 35 reliving sessions undertaken by 26 different participants. A total of 1,365 interactions were recorded in this (1,055 minute) process, which translated to slightly more than 1 click per minute. Hence the users liked to interact reasonably *actively* with the system. This is still much longer than typical browsing/strong interaction based reliving (e.g. Facebook), where we expect the mean time between clicks to be in order of seconds (e.g. ~7 seconds reported in [24] for a related task). Slide-shows, conversely provide no interaction mechanism, hence we would expect mean time between interactions to be in the order of minutes (e.g. 5 minutes).

We undertook detailed analysis across demographics based on the following questions:

1) Which demographic set interacts most actively with the reliving application, and which prefers to watch it passively?

The table shows clicks/minute recorded for different demographic groups.

Females	**1.14**	**1.49**	**1.13**	**1.01**
Males	**1.41**	**1.25**	**2.08**	**1.43**
Both	**1.30**	**1.27**	**1.28**	**1.35**
	All	**1st party**	**2nd party**	**3rd party**

As the above table indicates, we found that in general males tend to be more active than females. We also found that 3rd party users tend to interact most actively. The male and female behaviors however are quite different across their roles as first, second, or third parties.

Figure 12: Number of clicks (normalized to percentage) across different semantic axes.

Figure 13. 'Stickiness': Time spent viewing the show after click on different semantic axes

2) Which semantic axis is used most frequently by the users?

We also noticed that the most commonly used type of interaction was the skipping a few (one or more) events based on preview. This could imply that users have a 'surfing the channels' kind of mindset and perhaps get bored by seeing all the details (sometimes the event could contain up to 100 images) of an event. In fact in our user interviews, many suggested a more enhanced previewing and summarization support. This is a subject we will address in the near future. Our current attention is focused on the three semantic axes of people, places, and time (i.e. *'who, where, and when')* and how people interact with them in reliving settings.

Based on 747 clicks on different semantic axes, we found that the most frequently used axis for semantic interaction in the reliving sessions is 'People' (data shown as percentages in Fig. 12). 'People' was the most frequently used axis across (almost) all demographics. The only exception was 3rd party interaction in which 'Places' was the most frequently used criterion.

We also studied the stickiness i.e. 'the time spent viewing the show after a click' on the different semantic axes and found that the average time spent before changing semantic axis was 80.13 seconds. The relative time spent viewing the show after clicks (normalized as percentages) is shown in Fig. 13. Based on time spent, we found that 'Places' is most popular reliving criteria. It was the leader across all demographic groups except 1st person, where people tend to use 'Time' predominantly. Interestingly, we found no significant differences in male and female interaction patterns across different axes.

On the whole, we found that the use of different semantic axes peaked with different types of participants. *'Time' was the most frequently used by 1st party participants*, and *'People' are most frequently used in 2nd party interactions*. This makes intuitive sense as the primary participant is most likely to remember the time of occurrence of something they want to see. 2nd party users were basically family members hence their interest in 'People' also makes sense. Lastly, '*Places' are most commonly employed by 3rd party users.* They are typically least close to the people in the pictures and hence tend to be attracted to different 'Places' in the collection.

3) What are the common interaction patterns, and do they vary by demographics?

Lastly, we analyzed the interaction patterns i.e. 2nd order behavior such as which axis led to another in the user interaction-experience space. Transition graphs across the different semantic axes are shown Fig. 14. We noticed that more than half (51%) of interactions are loopback types (i.e. the same axis is used again). This indicates that users have a tendency to stay on one axis until something in another axis catches their attention.

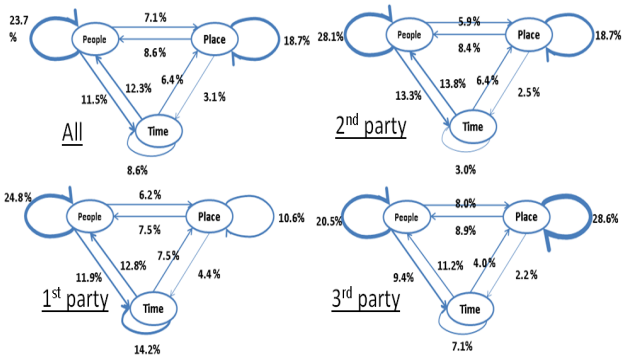

Figure 14. Transition graphs across the semantic axes.

We can again see the 'People' axis dominating and the 'Places' axis showing up prominently in the 3rd party case. We also notice that very few interaction patterns involve going from 'Places' to 'Time'. It is very small and in fact noticeably lesser than even its converse, i.e. 'Time" to 'Places' flow. This indicates that thinking of time first and then choosing across locations seems more common than thinking of place first and then choosing the time involved. On the other hand, the figure shows more pronounced movements between "People" and "Time" axes across all demographics somewhat indicating that users might be interested in viewing people across time.

6.3 Discussions

6.3.1 Reliving vs. retrieval vs. browsing

We would like to clearly differentiate between the problem focused "reliving on demand" with that of "media management". We consider media (image, video) as simply the portal into the user experience space. Media by itself is uninteresting unless it performs a function (e.g. reliving, sharing) for the human user.

Hence, while information retrieval is a valid problem in different contexts, it is clearly different from our focus. That said, a user can indeed search for certain media content along individual axis of time, location, people and event, or a combination thereof.

Similarly, while browsing has the potential for supporting 'reliving'; it typically occurs piece-meal, in an ad hoc manner. It does not create a holistic media show by combining multiple media elements in an aesthetically pleasing and multimedia enriched manner.

In a sense, reliving is well positioned between retrieval and browsing, and can be tailored towards either need to a large extent. We draw inspiration and support from the classification of users by clarity of their intent as "browsers", "surfers", and "searchers" in [6].

Our user study also indicates that there are different flavors of reliving with different viewers. We are pleased that such behavioral differences validate our design motivations and principles, at the core of which is to put the viewer in the driver's seat so that each individual can achieve a satisfying reliving experience.

6.3.2 Platforms

As we alluded in the introduction, digital reliving takes on all kinds of ubiquitous platforms (web, home, hardcopy) and devices (TV, PC, smart phones, tablets, digital picture frames, kiosks, photobooks). We believe the proposed media reliving system is suitable and adaptable for all of them. In fact, the selection and relative importance of different axes might become an important design consideration in devising both static (e.g. photo-book) and constrained real-estate (e.g. mobile) reliving scenarios.

6.3.3 Accuracy of media processing algorithms

The developed reliving system employs multiple multimedia processing techniques (viz. face detection, face recognition, event clustering, image geo-clustering, aesthetics value prediction) under the hood. These algorithms are not all 100% accurate. However the inaccuracies did not cause any observable disruption to the user experience (none of the 26 participants mentioned any issues with it). This can partially be attributed to the design choice made to focus on relatively robust semantic axes. The three axes of navigation are built on fairly reliable sources– 1) location from reliable clustering of geotags or (user given) folder names; 2) time from reliable camera metadata, 3) face information from reliable face clustering with user verification and labeling.

The underlying event clustering is based on reliable timestamps; and visual aesthetics is a soft factor that adds variety to the presentation. In summary, none of the potential errors would disrupt the user experience because the errors are benign (e.g. a missed face is often unnoticed). More importantly, the user attention is tightly geared to the far more engrossing reliving task rather than any single specific attribute or characteristic.

6.3.4 Relations to social media and cloud computing

We are currently witnessing an explosive growth in number of pictures and media elements being shared online across different media sites. Facebook (which has the strongest social aspect) has more picture uploads than any other site (around 2.4 billion new photo uploads each month). This clearly highlights the importance

of social sharing as pivotal to how media will be managed and used in the near future. Similarly, all this data is being made available in the cloud for anybody (with permission), anywhere to access it, thus empowering people to share their multimedia experiences with ease and enjoyment [14].

With all this proliferation and data explosion, the tools which allow users to engage with such media assets would become ever more relevant. We have adopted the approach of doing this based on the fundamental human purpose of media capture; that of reliving it. Our effort is focused on the postulate that the process of reliving would be different for each user with whom this data is shared. People will have different time availability, interest level and primary semantic axes across which they would like to relive these events. Realizing this, to the best of our knowledge, ours is the first attempt at providing the recipients with the flexibility and control to choose their axes, interests, and speed of engagement with the social media shared by others. We see this as providing the very valuable middle ground of user engagement between every-click-browsing and passive slide-show viewing. Our system transforms gracefully into those two extreme cases but also allows the interaction and reliving to occur on demand.

7. CONCLUSIONS AND FUTURE WORK

We have developed and evaluated a novel media reliving system that produces aesthetically appealing and semantically drivable multimedia slide-show based on reliving dimensions of events, people, locations, and time. We allow each viewer to interact with the default presentation to dynamically redirect on the fly the flow of reliving as desired from their individual perspectives. Furthermore, using the logged reliving sessions we have discovered many interesting findings on the reliving needs, behaviors and patterns of different users, validating our design motivations and principles. The patterns learned (e.g. varying importance of people, place, and time based on 1st, 2nd or 3rd party interactions) would provide guidelines for our future versions, as well as potentially for others designing reliving applications. Our studies have confirmed that liberty to navigate along the dimensions of "time, location, and people" and the freedom to change paths as and when desired are essential to *true reliving*, which cannot be currently realized even in state-of-the-art image viewing systems. We strongly believe that the proposed "Reliving on Demand" paradigm will redefine the way people view and experience multimedia in personal and networked environments.

There are a few major improvement opportunities. Users are interested in having the option of using multiple axes to perform search-like functions. The event thumbnails can be made more informative (e.g., mouse-over expansion to multiple samples of an event). Duplicates and duds can be excluded (at least with an option given that detection algorithms can make errors). It may be desirable to tag favorites while reliving (so one can playback all favorites at the end or they can given favorite treatment later on for the same or different viewers). Finally, more data mining and recommendation can be done to further enrich and empower media sharing.

Acknowledgments: We thank the numerous participants who contributed their pictures and time, and provided valuable input.

8. REFERENCES

[1] D. C. A. Bulterman. Using SMIL to encode interactive, peer-level multimedia annotations. *ACM Symposium on Document engineering, 2003.*

[2] L. Cao, J. Luo, A. Gallagher, X. Jin, J. Han, T. S. Huang. A Worldwide Tourism Recommendation System Based on Geotagged Web Photos. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2010.

[3] R. G. Cattelan, C. Teixeira, R. Goularte, and M. D. G. C. Pimentel. Watch-and-comment as a paradigm toward ubiquitous interactive video editing. *ACM Trans. Multimedia Comput. Commun. Appl.* 4(4), Article 28, 2008.

[4] J.-C. Chen, W.-T. Chu, J.-H. Kuo, C.-Y. Weng, and J.-L. Wu. Tiling Slideshow. *ACM Int. Conf. on Multimedia*, 2006.

[5] J. Chen, J. Xiao, and Y. Gao. iSlideShow: A Content-aware Slideshow System. *Int. Conf. on Intelligent User Interfaces*, 2010.

[6] R. Datta D. Joshi, J. Li and J.Z. Wang, Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys*, 40(2), 2008.

[7] R. Datta and J. Z. Wang. ACQUINE: Aesthetic Quality Inference Engine - Real-time Automatic Rating of Photo Aesthetics. ACM Int. Conf. on Multimedia Information Retrieval, 2010.

[8] J. Huber, J. Steimle, M. Mühlhäuser. Toward More Efficient User Interfaces for Mobile Video Browsing: An In-Depth Exploration of the Design Space. *ACM Int. Conf. on Multimedia*, 2010.

[9] A. Jaimes, D. Gatica-Perez, N. Sebe, and T. Huang. Human-Centered Computing: Toward a Human Revolution. *IEEE Computer*, 40(5), 30-34, 2007.

[10] R. Jain. EventWeb: Events and Experiences in Human Centered Computing. *IEEE Computer*, 2008.

[11] W. Jiang, A.C. Loui, C.D. Cerosaletti. Automatic Aesthetic Value Assessment in Photographic Images. *IEEE Int. Conf. on Multimedia and Expo,* 2010.

[12] Y. Ke, X. Tang, and F. Jing. The Design of High-Level Features for Photo Quality Assessment. *IEEE Int. Conf. on CVPR*, 2006.

[13] A. C. Loui, A. Savakis. Automated Event Clustering and Quality Screening of Consumer Pictures for Digital Albuming. *IEEE Trans. on Multimedia*, 5(3), 2003.

[14] W.-Y. Ma. Rethinking Multimedia Search in the "Clients + Cloud" Era. *ACM Int. Workshop on Large-scale Multimedia Retrieval and Mining*, 2009.

[15] G. Marchionini. Exploratory Search: From Finding to Understanding. ACM Communications, 49(4), 41-46, 2006.

[16] M. Rabbath, P. Sandhus P, S. Boll. Automatic Creation of Photo Books from Stories in Social Media, *ACM Int. Workshop on Social Media*, 2010.

[17] K. Ren, R. Sarvas, and J. Ćalić. Interactive Search and Browsing Interface for Large-scale Visual Repositories, *Multimedia Tools and Applications*, 2010.

[18] Y. Rui, T.S. Huang, and S. Mehrotra. Content-Based Image Retrieval with Relevance Feedback in MARS. *IEEE Int. Conf. Image Processing*, 1997.

[19] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(12), 1349-1380, 2000

[20] C.G.M. Snoek, K.E.A. van de Sande, O. de Rooij, B. Huurnink, E. Gavves, D. Odijk, M. de Rijke, Th. Gevers, M. Worring, D.C. Koelma, A.W.M. Smeulders. The MediaMill TRECVID 2010 Semantic Video Search Engine. *NIST TRECVID Workshop*, 2010.

[21] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2001.

[22] J. Xiao, X. Zhang, P. Cheatle, Y. Gao, and C. B. Atkins. Mixed-Initiative Photo Collage Authoring. *ACM Int. Conf. on Multimedia*, 2008.

[23] S. Xu, T. Jin, F. C. M. Lau. Automatic Generation of Music Slide Show Using Personal Photos. *IEEE Int. Symposium on Multimedia*, 2008.

[24] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted Metadata for Image Search and Browsing. *ACM SIGCHI Conf. on Human factors in Computing Systems*, 2003.