Multimed Tools Appl (2012) 56:9–34
DOI 10.1007/s11042-010-0538-7

Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications

Mor Naaman

Published online: 21 May 2010

Abstract In recent years, various Web-based sharing and community services such as Flickr and YouTube have made a vast and rapidly growing amount of multimedia content available online. Uploaded by individual participants, content in these immense pools of content is accompanied by varied types of metadata, such as social network data or descriptive textual information. These collections present, at once, new challenges and exciting opportunities for multimedia research. This article presents an approach for "social multimedia" applications. The approach is based on the experience of building a number of successful applications that are based on mining multimedia content analysis in social multimedia context.

Keywords Social media · Flickr · Youtube · Multimedia · HCI · Data mining · Multimedia applications · Evaluation

1 Introduction

We live in an era of change for multimedia research and applications. The ease of multimedia content production, coupled with exceedingly lower cost of publishing and wide potential reach, result in a staggering amount of content available on the Web. Social Media services and web sites such as YouTube [107] and Flickr [35] allow people to share this multimedia content in an immense scale. To illustrate, Flickr users have shared over 4 billion images and videos on the site as of November 2009[1] and Facebook users share a similar amount of photos each *month*.[2] It was also

[1] http://blog.flickr.net/en/2009/10/12/4000000000/

[2] http://www.facebook.com/press/info.php?statistics, retrieved March 2010.

M. Naaman (✉)
School of Communication and Information, Rutgers University, New Brunswick, NJ, USA
e-mail: mor@rutgers.edu

reported in 2009 that YouTube users share 20 new hours of video content every *minute*.[3]

This article proposes the term *social multimedia* to refer to multimedia resources available via social media channels, or more formally: *online sources of multimedia content posted in settings that foster significant individual participation and that promote community curation, discussion and re-use of content.* Social multimedia presents a significant opportunity for Multimedia applications and services [13]. Beyond the scale of available content, such services make new context information and metadata about the content widely available. Such information may include many facets: textual descriptors, information about the location of the content capture [57, 97], the camera properties metadata, and even user information and social network data. These additional metadata can be used to advance and augment multimedia and content analysis techniques. In addition, social multimedia captures and leverages community activity around multimedia data, using explicit user input like tags and comments (e.g., [23, 82]) as well as implicit input from users like mass viewing patterns in item and sub-item levels [79]. Indeed, social multimedia also offers the opportunity to design *interactive* systems that elicit new explicit and implicit metadata from user interaction. Such interaction and user input is often driven by social motivations [6, 65] and can improve the data available for multimedia applications. Thus, social multimedia offers several opportunities that go beyond and above other "Web multimedia" sources where many of these opportunities are not available.

Regardless of data source and scale, multimedia content analysis is still a difficult problem. Famously, the semantic gap was defined by Smeulders et al. [90] as the discrepancy between the information that one can extract from the visual data and the interpretation that the same data holds for a user in a given situation. Even recent advancements in computer vision (e.g., [51]) do not seem to make the semantic gap problem anywhere closer to being solved. Thus, many open problems in multimedia cannot yet make satisfactory use of content analysis techniques alone.

At the same time, social media is by no means a magic pill, especially considering that it is not free of its own significant limits and challenges. The aforementioned context and available metadata are noisy and often inaccurate, wrong or misleading [18]. As a result, there is very little "ground truth" for social media data. The noise and lack of semantics make even the simplest of metadata, user-provided tags, difficult to use. For example, a single video tagged `Bay Bridge` does not disclose to us which Bay Bridge the tag refers to, and it might not even depict any Bay Bridge in it whatsoever. Further, the lack of semantics means that there is no "right and wrong" in tagging: that `Bay Bridge` video may have been captured on a trip to see the said Bay Bridge; or maybe taken from the bridge but does not show the bridge itself (in both cases, the tag still carries some useful meaning for the user who assigned it). These issues of accuracy rise even before we consider issues of Spam and malicious content that add further challenges in open, public systems.

Importantly, social multimedia search and mining entails shift of focus from traditional multimedia applications. The availability of content does not require general detection and classification tasks (like, say, identifying "tigers" or "architecture"). Instead, tasks that are narrower in scope are emerging (for example,

[3] http://youtube-global.blogspot.com/2009/05/zoinks-20-hours-of-video-uploaded-every_20.html

"identify content from last night's U2 concert"). Second, and related, these new multimedia tasks and applications are often not driven by recall, but by precision, representativeness, diversity and effective presentation. In other words, for many applications, it is not important to retrieve *all* relevant social multimedia resources. Instead, identifying relevant resources in a highly precise manner might be more beneficial (e.g., finding a few representative photos of landmarks).

This article describes two specific applications that are built on mining of multimedia data. The applications represent the experiences of the Yahoo! Research Berkeley team, and the team's efforts in multimedia research over a number of years (e.g., [4, 5, 39, 45, 46, 48] and others). The applications could help illustrate some of the new opportunities embodied in social multimedia. The article generalizes our approach for the two applications, to suggest a general approach for social multimedia analysis and applications. Note that this article does not focus on computational models, heavily researched and discussed elsewhere (e.g., [17, 44, 104]). Instead, the focus of this work is on the unique properties of social multimedia and the new search and mining applications it enables. The ideas described in this article, then, can be used as conceptual guidelines for developing new real-world applications and services of social multimedia.

The generalized approach to social multimedia applications is described here as a series of steps, including:

Step 1: Identify topic and application domain and use simple context-based tools to identify relevant content items.
Step 2: Use application-specific, constrained and "knowledge-free" (unsupervised) content analysis techniques to improve precision, representation and selection of items.
Step 3: Use the content analysis output to further improve metadata for aggregate multimedia items.
Step 4: Leverage user interaction for improving relevance and representation.

This article lays out the ideas behind the four steps outlined above, illustrated using the two applications. Although the applications have different goals and use cases, they both rely heavily on social media sources and data as well as multimedia content analysis. The purpose of this article is not to elaborate on the details of each application, but rather focus on the principals and the unifying concepts that played a role in both. To this end, the text is organized according to the four steps listed above. It begins, however, with a broad but brief summary of related work in the area of social multimedia, followed by a short introduction to the two sample applications discussed in this article. A closing section discusses considerations for the evaluation of social multimedia applications.

2 Related work

This section considers related research in various areas, beginning with the general theme of social multimedia, including work on multimedia and social tagging which helped expand the scope of multimedia applications over the "early years" of content-based analysis [29, 49, 50, 90]. This discussion is included to better situate

our two applications with the other research in the field. The section then describes specific efforts that are directly related to the two applications featured in this article.

2.1 Social multimedia

As defined above, social multimedia offers different avenues for research in the multimedia domain, including: analyzing community activity around multimedia resources; deriving metadata from social activity and resources; and pooling of content in social settings (the latter is discussed in relation to our Concert Sync application in Section 2.3).

One potential benefit of social multimedia is the opportunity to aggregate data or analyze activities around individual resources to better reason about their content. For example, Shamma et al. [79] use chat activity in Instant Messenger to reason about the content of shared online videos; De Choudhury et al. [23] analyze comments on YouTube videos to derive "interestingness" and topics; and Mertens et al. use community activity for "social navigation" of web lectures [54]. In a more recent work, Shamma et al. [81, 82] use the content, volume and trends of Twitter messages about a multimedia broadcast (e.g., the US presidential debates of 2008) to reason about the content of the event. In the work presented here, we do not aggregate activity around a single resource, but rather use different methods to pool content items together for the analysis. Yet, the analysis we use for the pooled content is related to these efforts mentioned above, and is reflected in steps 1 and 3.

Indeed, beyond individual items, aggregate trends and data could be derived for *multiple* content items or for the entire collection to help in visualizing or browsing a collection. Such work includes, for example, visualizing Flickr tags over time [32], reasoning about Flickr groups [63, 64], or extracting semantics of multimedia tags [75] and the relationships between them [77, 105]. Researchers also extracted location multimedia summaries and travel suggestions from aggregate social media blog and image data [38, 41, 109]. For example, Jing et al. proposed an algorithm that uses a text-based approach to extract representative sights for a city [42], and propose a search and exploration interface. Community activities were also used to augment and improve metadata about multimedia resources, like generating and displaying tag suggestions [55, 87] and augmentation of personal content using social media sources [16, 34].

In this work, one of the main aspects of social multimedia is the additional information the social media "context" adds to multimedia tools and applications and that enables improved content analysis. The topic of context augmentation of content analysis in multimedia research has been relatively active and widely discussed [14, 40] in the last 5–7 years. For example, a number of efforts used camera settings and/or capture time [28, 52, 89, 104] together with content analysis to improve performance of automated content analysis. Location context, "geotagged" or "geo-referenced" metadata, had played a major role in multimedia research since 2003 [30, 43, 60, 97]. Context metadata had also played a major role in personal multimedia management systems, utilizing metadata such as capture time and location as well as other sources of context (e.g., social) to organize photo collections [2, 15, 30, 60, 69, 73] and improve content analysis [19]. Mostly in the domain of personal collections and family photo albums, researchers employed context metadata to help face recognition [31, 56, 66]. This article does not propose

a new approach to merging context and content analysis, but rather highlights the opportunities in leveraging the context data available for social multimedia search and mining.

Finally, researchers have been using the sheer volume, as well as the unique metadata (e.g., tags) associated with social multimedia content as a resource for "traditional" multimedia tasks such as improving visual models and training concept detectors. For example, researchers performed learning of distance metric using tagged images [74, 106], and tried to infer and learn concepts [78, 95] and visual words [102] from "noisy" tags.

2.2 Work related to Flickr Landmarks

This section reports briefly on work related to our first application, Flickr Landmarks(see below). The focus of this section is mostly computer-vision approaches to "landmark" applications, and systems that aim to identify representative images. For a more detailed discussion, see [45].

Most closely related to the work here is the research from Simon et al. [88] on finding a set of canonical views to summarize a visual "scene" based on unsupervised learning using the images' visual properties (most prominently, SIFT features [51]). That work partially follows the approach we suggest here, although the authors do not employ an initial, context-based filtering step, leaving their description somewhat incomplete (see also [91]), not specifying how initial sets of content will be automatically generated under their scheme. Recently, Crandall et al. [25] have extended our landmarks work [45] by providing a more scalable solution for the landmark identification task. Indeed, the authors use ideas that are in line to the work described here, e.g., a specific task and sub-domain and use of application-specific properties (e.g., expected size of location clusters) to guide unsupervised learning. A similar approach is taken by Chen et al. [22]. Another attempt at scaling the landmark recognition is provided by Zheng et al. [108]. Others landmark-related efforts have used tag data [1] to classify photos based on their likelihood to be of a landmark, and other sources like travel guides (in addition to geotagged photos) [109] to identify landmarks.

Earlier work on topics related to landmark recognition was mostly applied to limited or synthetic datasets. Various efforts [11, 21, 31, 59, 67, 68, 98] examined different aspects of the problem, often performing analysis of context metadata together with content-based techniques. Slightly different approach was used in [47], where the authors investigated the use of "search-based models" for detecting landmarks in photographs, focusing on the use of text-based keyword searches over web image collections to gather training data to learn models.

2.3 Work related to Concert Sync

This section reports briefly on work related to the second application described here, Concert Sync. The focus of the section is in a number of areas, including: event-based management of media, research on video summarization, and work on multimedia related to live music or concerts. For a more detailed discussion, see [46].

Work on event-based management of media mostly considered personal events in stand-alone, personal systems (e.g., [36, 60] and more). Some research had explored

the mobile, situated experience of event multimedia [76]. Lately, the event construct was expanded to include social web-based representation [110] and other aspects of event modeling [103].

Projects related to video summarization and selection of key-frames were mostly based on content analysis [24, 99], but some community-based methods have recently been proposed, like using community "remix" data for summarization tasks [83] or using viewing activity [79, 94] to reason about the video content. The model and application scenario presented here are quite different than all of the above, yet can potentially be used for similar tasks.

The domain of media from live music events was the focus of several research projects [61, 92, 100]. These efforts mostly looked at ways to present professionally produced or "authoritative" video or audio content (e.g., a complete video capture provided by the event organizers).

The contribution from Shrestha et al. is most related to the core technical aspects of the Concert Sync work, looking at synchronizing and aligning multiple video clips of a single scene. The authors' first approach involved detection of camera flashes in video clips from multiple contributors [85]. Later, the authors used audio fingerprints (much like we did in our work) for content synchronization [86]. The work described in this article shifts the focus of the system from developing matching algorithms to mining the structure of the discovered overlaps and audio re-use to create compelling new ways of aggregating and organizing community-contributed Web data.

3 Sample applications

This article builds on our experience with two social multimedia applications: Flickr Landmarks and Concert Sync. The applications draw from different content types (images and video), content sources (Flickr and YouTube), and have entirely different goals and use cases. At the same time, both applications demonstrate the concepts and the different steps in search and mining, as well as presentation and evaluation, of social multimedia applications. The two applications, as reported below, were implemented in parts; notes in the respective sections indicate ideas that are not yet implemented.

The idea behind Flickr Landmarks [45, 48] is to improve the task of searching and browsing for photos of local landmarks. Sample user scenarios include search and tourism exploration [5]. The system mines social media data (namely, Flickr photo metadata) to automatically generate a list of terms that are likely to represent key landmarks and attractions in any region, worldwide. Then, the system can automatically select diverse and representative photos of these landmarks. As a result, a user can quickly search or get an overview of what are the landmarks and attractions in a given area, as well as get a good visual representation of each landmark. Figure 1 presents a browsing interface for the Flickr Landmarks system. As reported in this article, we implemented the browsing and exploration part of this application, and had shown independently (in slightly limited settings) how to filter the selected explored concepts to only include landmark-related concepts, and how to use content analysis to improve relevance and representation.

The second application, Concert Sync, aims to improve the representation of videos captured at the same live music event. The availability of video capture

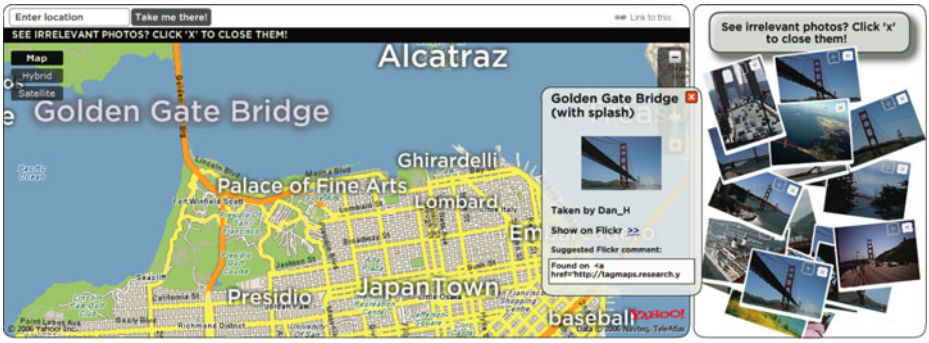

Fig. 1 A screenshot from the World Explorer visualization, showing parts of San Francisco; the user highlighted the tag Golden Gate Bridge to bring up photos with that tag from that area, and then selected one of the photos to get an expanded view

devices, and the high reach and impact of social video sharing sites like YouTube [107], make video content from live shows relatively easy to share and find [27]. However, there are new challenges that impede this new potential: issues of relevance, findability, and redundancy of content, even from a single event. Our system automatically mines YouTube to aggregate, organize and add metadata to videos taken at the same music event. The result is a much-improved user experience when searching, viewing or browsing content from that event. Figure 2 presents a possible browsing interface of Concert Sync. As detailed below, we implemented the different portions of the system and tested them on content from several different concerts. We have not released a complete application that includes the user experience described below, but we do provide ideas regarding the new types of experiences our system enables.

The next sections provide the details on the common generalized steps for the two applications. These steps are by no means a strict prescription for social multimedia applications. The steps could be used, though, as practical guidelines to help social multimedia researchers conceptualize and develop new services and systems. Some

Fig. 2 A possible interface for browsing concert video clips by synchronizing playback

ideas for new systems and services that follow this approach are provided in the conclusion of this work.

4 [Step 1] Using context to identify relevant content

The first task at hand is to reduce the amount of content to be examined and analyzed. With billions of pieces of content available, the task of content match can be difficult and unreliable, regardless of method. Firstly, even the initial task of extracting content-based features could prove to be a significant challenge at this scale [20]. Beyond feature extraction, content indexing and matching is also problematic in a large repository, especially in a high-dimensional search space. Content match methods like nearest neighbor [33], for example, are often unreliable [12]. In addition, the potential use of supervised learning techniques is limited: the "long tail" of resource categories (e.g., different tags on Flickr) does not allow training for each individual concept that may appear in the data.

Luckily, social multimedia offers a plethora of context information that can be used to filter content items. Such context information includes, for example: text associated with the content (title, description, tags, comments), location and time metadata, personal and social data (including "social"/contact network), viewing data (including view count and other view metadata such as scrub, stop and pause actions for video [79]), and capture device metadata (e.g., camera properties data available from a photo's EXIF header). Note that our sample applications do not use all the context dimensions listed above. For example, neither application makes use of the contact network information on Flickr or YouTube, although scenarios where such data might be applicable can be easily derived for both applications.

The following subsections describe how we, for Flickr Landmarks, identified tags that are likely representative of landmarks and retrieved the matching content for each; and for Concert Sync, identified YouTube videos that are likely captured at a given live music show.

Note that these tasks represent the fact that, in both applications, we identified both the relevant resources, and the specific *sub-topics* they match, based on the context information. In other words, our bottoms-up approach did not only identify photos that are related to landmark, but did it in a way that identifies and groups the photos belonging to each unique landmark (and, respectively, music concert). Such content is thus highly relevant to a single concept in the domain of choice. In other words, these step will result not only in a set of content items S that are relevant to the task at hand, but also in a set of clusters C_S that group together related content items.

4.1 [Flickr Landmarks] Using context to retrieve landmark photos

To create a dataset of landmarks and their photographs, we first identify a set of landmarks using the metadata of the Flickr items. This section briefly describes our approaches for extracting the tags that represent geographic features or landmarks. These tags represent highly local elements (i.e., have smaller scope than a city) and are not time-dependent. Examples may be `Taj Mahal`, `Logan Airport` and `Notre Dame`; counter examples would be `Chicago` (geographically specific but

not highly localized), `New York Marathon` (representing an event that occurs in a specific time) and `party` (does not represent any specific event or location). While this is quite a loose definition of the concept of landmark tag, in practice we show that our approach can reasonably detect tags that are expected to answer these criteria.

The approach for extracting landmark tags using context information is two-fold. First, we identify representative tags for different locations inside an area G. In the second part, we check whether these tags are indeed location-specific within area G, and that they do not represent time-based features.

The first part of the process is described in detail in [5], and consists of a geographic clustering step followed by a scoring step for each tag in each cluster. The scoring algorithm is based on TF-IDF, identifying tags that are frequent in some clusters and infrequent elsewhere in the same (city-scale) geographic area. The output of this step is a set of high-scoring tags x and the set of location clusters L_x where these tags occur. Thus, these techniques can detect geographic feature tags as well as the locations where these tags are relevant, given a geographic region as input. For example, in the San Francisco region, this system identifies the tags `Golden Gate Bridge`, `Alcatraz`, `Japan Town`, `City Hall` and so forth.

The second part of our proposed landmark identification is identifying individual tags as location-driven, event-driven or neither. We can then use the already-filtered list of tags and their score (from the first part of the computation), and verify that these tags are indeed location-driven, and that the tags do not represent events. The approach for identifying these tag semantics is based on the tag's metadata patterns; the system examines the location coordinates of all photos associated with x, and the timestamps of these photos. The methods are described in more detail in [75]. For example, examining the location and time distribution for the tag `Hardly Strictly Bluegrass` (an annual festival in San Francisco), the system may decide that the tag is indeed location-specific, but that the tag also represents an event, based on its temporal distribution.

The output of this process is a set of tags x and a set of locations L_x where each tag is relevant. From the set of tags and locations we can further get relevant photos from Flickr: photos with the given tag, taken around the respective location cluster. This set of tags and groups of photos could now be used in content analysis tasks. As we show in [48], this set already exhibits higher precision than photos retrieved just by using the landmark tag, without the location metadata.

4.2 [Concert Sync] Using context to retrieve concert videos

To retrieve video clips taken at a specific event, we start with structured information about music concerts that had taken place. There are a number of potential resources for such structured listings: the social event sites SonicLiving and Upcoming.org, as well as Facebook and Last.fm, are examples of sources for concert information. All these websites feature event listings for music events. Listings usually include the name of the performing artist/band (or bands), the name and details of the venue, and the date and time of the show. While the focus of these sites is forward-looking, some of these sources also make historic (past) data about concerts available.

From an event listing on one of the event websites, we can construct a set of queries to YouTube, Flickr or other social media content services. For our initial implementation, the query construction was performed according to rules, based

on a number of heuristics. For example, an initial query could include the name of the band and the venue, and retrieve content that was uploaded on the date of the concert, or within days after the concert took place (to avoid general content that is not related to the concert but is related to the band). Another query can use the geographic and time metadata of the concert, together with the band name (e.g., videos with the text "Calexico" taken on Sep 24th, 2008 in New York).

Some social media sources may necessitate turning to alternative resources to improve the context information used for these queries. The metadata needed for some of these queries is not always available, and the content sources often do not allow queries of certain type. For example, videos captures often do not include capture time metadata, and YouTube does not currently allow search by arbitrary date range. To overcome some of these limitations, we can use other services as sources for additional context that can be used to improve YouTube queries. One strategy is to first mine information from Flickr, using the richer metadata and API available there. Once content from Flickr was retrieved, the content's metadata can be used for mining YouTube content. For example, we can consider the set of tags associated with the band-location-time query, and compare those tags to the tags associated with photos retrieved by using the only-band-name query. Popular tags that appear only for the more-specific query are used to generate a query to YouTube. Those tags, we found, often include the name of the city where the event took place, or alternative names for the venue.

Alternatively, in our most recent work [8, 9], we used an approach that does not require a-priory knowledge of event listings. Instead, we used a clustering-based approach to identify content clusters, where clusters represent content captured in the same event. For the clustering, we exploit the rich context associated with the content, including textual annotations (e.g., title, tags) and other metadata such as content location and creation time. We defined appropriate document similarity metrics, exploring a variety of techniques for learning multi-feature similarity metrics in a principled manner. Training data was used to inform a clustering framework. While we performed the evaluation on a large-scale dataset from Flickr, similar ideas could apply for identifying event content from YouTube.

The end result of this process is a curated set of content, including video clips that were likely captured at the same live music event. Notice that for this application, as well as for Flickr Landmarks, the process of getting relevant content focuses on precision, not on recall. In other words, false negatives are permitted, as the applications do not demand a complete coverage, as long as there is an extensive set of resources for every concept (as is indeed the case for events or landmarks that are popular enough, which is increasingly the case in this age of abundant content). On the other hand, 100% precision is not required either. The content-based methods will help in handling the false positives, so that those can be rejected later.

5 [Step 2] Using robust, application-specific content analysis

Once we have used context information to gather relevant content resources matching the specific concepts of interest (events, landmarks), we can employ content analysis methods to improve the representation and organization of the collection. The idea is that given both a constrained application domain, and an already-filtered

set of resources, the content-based techniques could be applied in a robust and useful manner. We can apply visual methods that are specific to the content features that are key for an application, instead of simply using a bag of visual features that may or may not be relevant. In addition, in our work we opted for using unsupervised content analysis techniques since, due to the long tail of content, one cannot hope to train classifiers on all the concepts that will appear in the dataset.

The rest of this section provides an overview of the content analysis tasks and techniques we used for Flickr Landmarks and Concert Sync. In the first, we used the fact that the application (landmarks) suggests the existence of a dominant feature that would appear in most photos and could be detected by using geometric features. In the second, we used the fact that the video clips were all likely to capture segments of the same audio source.

5.1 [Flickr Landmarks] Finding representative views

Given a set of images that are likely to represent a specific landmark, we can employ algorithms that leverage the expectation of similar landscape and a common object that appear in most images. Thus, our constrained domain made the content analysis possible and, ultimately, robust. On top of the constrained domain, the reduced set of images that resulted from earlier context-based steps made the content-based process feasible (limited number of resources to analyze) as well as more accurate (less noise and fewer false positives).

Figure 3 shows the outline of the content analysis process. Given the reduced set of images, our approach for generating a diverse and representative set of images for a landmark is based on identifying "canonical views" [72, 88]. We use image analysis to cluster the landmark images into visually similar groups, as well as generate links between those images that contain the same visual objects. Based on the clustering and on the generated link structure, we identify canonical views, as well as select the top representative images for each such view.

To this end, we extracted both global and local features of the images. The global features we extracted are grid color moment features [93], representing the spatial color distributions in the images, and Gabor textures [53], representing the texture. As for the local features, we extracted local interest point descriptors modeled and represented via the scale-invariant feature transform (SIFT) [51]. First, we used the global features (color and texture) to discover clusters of similar images within a given set of photos for a single landmark. Second, we ranked the clusters using their internal visual coherence and likelihood to represent the same object (using the fact

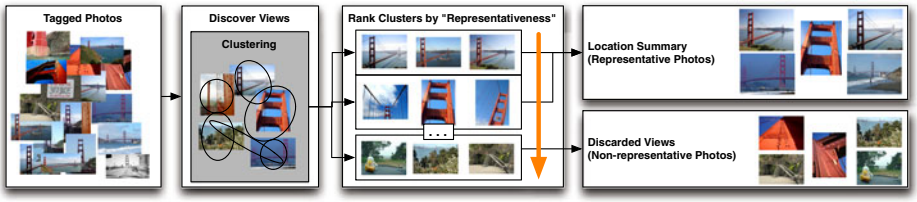

Fig. 3 Illustration of the process for generating representative summaries of landmark image sets

that we were expecting a visual landmark in the photos). We used various features including SIFT for this step.

Finally, we used similar methods to rank photos in each individual cluster according to how well they represent the cluster. This analysis was based on the assumption that (1) representative images will be highly similar to other images in the cluster, (2) representative images will be highly dissimilar to random images outside the cluster, and (3) representative images will feature commonly photographed local structures from within the set. For a detailed description of the content methods, as well as other factors that we used in the ranking of the clusters and photos, see [45].

The end result of this content-based process is a further breakdown and organization of the content for each landmark. For a given landmark, top-ranked clusters represent different views, and the top-ranked images for each cluster can be shown in search results or otherwise when showing images for that landmark. Section 6 shows, in Step 3, how this structure can be further exploited to improve the metadata and information about the landmark.

5.2 [Concert Sync] Synchronizing video clips

Given a set of video clips that were likely taken at a specific music event, we can employ algorithms that leverage the expectation of the same audio source to look for overlapping segments amongst the different clips. Notice that using audio to identify overlap of arbitrary video clips is probably not an easily tractable problem: the vast amount of content and the low quality of recording may make the feature extraction prohibitive, and the indexing and matching too difficult. However, two factors that we built on allowed for a successful overlap identification. First, we made use of the fact that our clips audio content is music, and not speech or other sounds. Second, we used the context to identify a relatively small number of resources to analyze and compare. We could therefore reliably find overlapping segments amongst our concert resources.

Figure 4 illustrates the processing steps executed by Concert Sync. We used Audio Fingerprinting [37, 101] to synchronize the content clips captured by users at a certain show. In other words, we used the clips' audio tracks to detect when same the moment is captured in two different videos, identify the overlap, and specify the time offset between any pair of overlapping clips. We applied the method proposed by Wang [101]. Briefly, the approach operates by taking the short-time Fourier

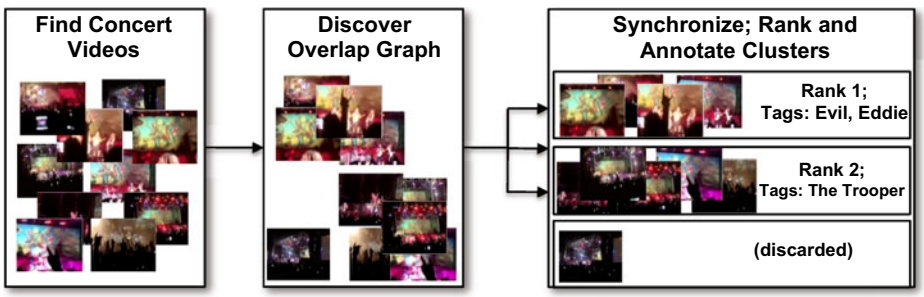

Fig. 4 Illustration of the computation process for Concert Sync

transform of a given audio segment and identifying "landmarks," which are defined to be the onsets of local frequency peaks. Each such landmark has a frequency and time value. The fingerprints associated with the segment are determined by constructing hash values for a set of target landmarks using the time and frequency differences between the landmark and a few adjacent landmarks. The result of the fingerprinting process is a large set of time-stamped fingerprints, where the fingerprints are simply hash values.

The task, then, is to determine whether or not any two given audio clips are recordings of the same audio source. This detection task is performed by finding all occurrences of matching hash values between the two clips. Two matching clips will have a great proportion of the matching hash values occurring at identical offsets in each of the two clips. The detection of a match between two clips is thus reduced to detecting the presence of a unique offset between the two clips that contains a sequence of many matching hash values. As we report in [46], we can set the parameters such that near-perfect precision for our dataset is maintained at a pairwise-recall level of 20%–30%. This level is sufficient given the specific application.

The end result of the content-based analysis is a synchronized set of audio segments and links between them. The synchronization of clips enables a novel experience for watching the content from the event, improving the user experience and reducing the redundancy of watching multiple clips of the same moment. Figure 2 presents one possible viewing interface. The figure suggests that the playback of overlapping clips is synchronized. The seven clips displayed all advance in unity, showing the same moment of the show. Clicking any one of the clip frames will switch the selected video into the frame of the right-hand side of the interface, showing the selected video in more detail. We therefore allow the user to select the preferred angle, viewpoint or the best-quality video for any given moment of the show. Once a clip's playback ends, that clip would fade off the screen. New clips that overlap with the current timeline would automatically appear during playback. Beyond synchronized playback, the synchronization and overlap data help improve both findability and relevance of clips from the event, as shown in the next section, followed by a discussion on how the user's clip-viewing selections could further improve the content metadata.

6 [Step 3] Content match improves metadata

The links between content resources that were extracted using the content-based methods can be used to further improve the metadata and organization of the aggregate content. The critical element behind the ideas described in this section is that aggregate patterns, which could not be exposed before, are now available based on the clustering or grouping of content items. These patterns can be use to enhance or better understand the metadata and context of capture.

This section demonstrates the different treatments we applied to the content analysis output in both applications to produce additional metadata that could help with content organization and retrieval. For Flickr Landmarks, the potential techniques listed were not implemented, and are provided here simply to illustrate the generality of this step. For Concert Sync, we had implemented and tested three

different techniques, including using the content matches to get descriptive text for browsing, to determine audio quality, and to rank moments from the live show according to their "importance".

6.1 [Flickr Landmarks] Finding viewpoints and descriptors

Having identified the visual groupings of the photos of a certain landmark, we can examine the aggregate patterns to extract additional metadata about those groups. The description below hypothesizes on the types of information that could be extracted. It is important to note that these ideas were *not* implemented and are included here for completeness.

At least two types of metadata can be improved based on the content match in Flickr Landmarks: textual descriptors and aggregate location metadata. Our unit of analysis here is all the photos of a given landmark as identified in Step 1, and the clusters resulting from the analysis of these photos in Step 2. Comparing the textual tokens (based on tags, descriptions, titles) of the photos in a single cluster to tokens associated with photos in other clusters, we are likely to find terms that are popular in one cluster and not in others. Such terms can potentially identify the unique aspects of the different views or viewpoints of the given landmark. For example, we expect one of the representative views of the Bay Bridge in San Francisco to include photos taken from the *Ferry Building*; this term should be represented more heavily in the appropriate visual cluster. Other potentially significant terms can be, perhaps, more visual in nature, like "panorama", or "interior". Such terms can improve the presentation of the different groups of photos, whether in search or visualization task.

Second, the visual groups created by the content match can potentially help in refining the location metadata of individual photos in the group, and possibly even help in generating orientation metadata. The visual clusters are not directly influenced by the location information, but at the same time, are expected to include photos taken from roughly the same location and angle. The location metadata for individual photos is not always accurate or correct; however, if we detect a cluster whose photos represent a relatively narrow area (or location trajectory—most of the photos are on a single line, for example), it might be possible to add, refine or correct the location and orientation information for some of the photos in that cluster that do not match that area or trajectory [91].

Again, we did not implement or test these hypotheses, and they are shown here as examples for applying the ideas of Step 3 to the Flickr Landmarks application. The next section discusses the implementation of similar ideas to improve and enhance the Concert Sync system.

6.2 [Concert Sync] Extracting interest levels, descriptive terms and audio quality

How does the content match improve available metadata for Concert Sync? Once synchronized, we used both the relative time information and links between overlapping clips in Concert Sync to generate important metadata about the clips and the event. The enhanced metadata can offer better representation and information in browsing collections of event videos. This metadata is derived from the collective video recordings and sharing behaviors of multiple event attendees, and made available by the content analysis step.

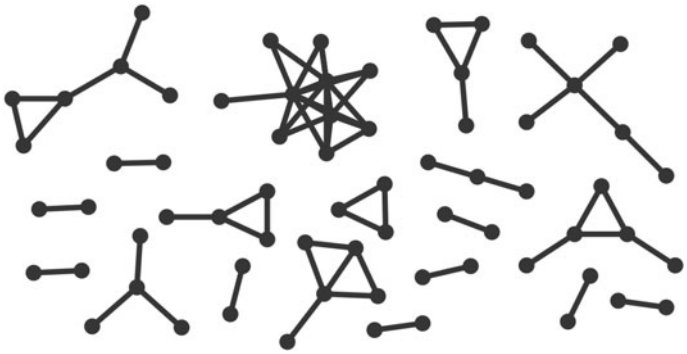

Fig. 5 Example graph structure emerging from linking overlapping video clips from one concert. The large connected component might suggest an "important" moment in the show

The key to extracting the different metadata is the observation that the emerging graph structure of matches between clips can lead to rich cues about the content of video clips and the event. The implicit graph structure is generated by creating a link between each two clips that were found to overlap. Such a graph from a real-life event is shown in Fig. 5. For example, the top-left cluster in the figure shows six clips, and six matches generated between those clips. Two of the clips in the figure were detected as overlapping with only one other clip, although it is *possible* that other overlaps exists in this cluster or beyond but were not detected.[4]

The various methods we applied to generate metadata using the link structure are discussed next: we identified level of interest [62] and significant moments in the show; mined the tags of overlapping videos to extract semantically-meaningful descriptive terms for the key moments in the show; and found the highest-quality audio recording of any time segment, given multiple overlapping recordings.

First, we used the graph to identify "important" moments in the show. We hypothesized that the segments of concerts that are recorded by more people might be of greater appeal to content consumers. Identifying these segments can be helpful for search, summarization, keyframe selection [24] or for simplifying the exploration of the event media. One possible interface would highlight clips captured during the most important segments and at the same time filter low-scoring clips that are either unrelated to the concert or (presumably) less interesting. Our system assumed that larger clusters of matches between clips correspond to segments of the concert that are subjectively most "interesting." This assumption was translated into a very simple measure of ranking importance of clusters: simply counting the number of nodes (video clips) each cluster contains.

Second, we used the graph structure and the user annotation for each individual clip to extract key terms that can represent or explain the content in individual or overlapping clips, and thus better describe and capture the overall themes that appear in the video clips. For this task, the system incorporated the lightweight annotations assigned to the media by the users in the form of titles, descriptions,

[4] A connected component of this graph does not necessarily mean that all the connected clips actually overlap: the overlap property is not transitive.

or tags for each video. Intuitively, we expect the overlapping videos within our discovered clusters to be related, and therefore expect the users to choose similar terms to annotate these videos – such as the name of the song being captured or a description of the actions on stage. We can identify terms that are frequently used as labels within a given cluster, but used relatively rarely outside the cluster. We used a simple scoring method based on TF-IDF, where the clusters generated by the content analysis step serve as the documents. The details of the computation appear in [46]. Indeed, our evaluation demonstrated that this technique could provide reliable, descriptive tags for a set of videos. For example, extracted terms in many cases included song titles, such as "Intervention," "Laika," "Hallowed Be Thy Name," "Fear of the Dark," and "Da Funk" for clusters of videos from an Iron Maiden show.

Lastly, the content match can provide a measure of the audio quality of individual clips, an important (yet not readily available) metadata. The quality of audio could prove to be rather important for the end-user application (i.e., synchronized playback of clips): inevitably, given the nature of user-generated recordings, the video and audio quality and content can be highly variant between clips, and the user may benefit by the system playing the best-quality audio for any segment of the show. The video playback, whatever it may be, can be overlaid on top of the automatically selected highest-quality audio track.

To find the highest quality tracks for each cluster, we use the fact that higher quality recordings are more likely to match other recordings in the dataset. We choose the most-connected video clips in each cluster as the probable highest-quality audio tracks. Note that automated content-based methods to extract audio quality (such as PEAQ [96], PAQM [10], and PERCEVAL [71]) require original source audio, which is not available here, and are optimized for verifying codec performance. Instead, we utilize the simple, already-available content match from previous step to reason about audio quality. A more comprehensive solution could combine this method with content-based audio quality metrics for improved results.

More details on these metadata-enhancing methods, as well as an elaborate evaluation of each method on a real-world dataset, are available in [46]. Step 4, ahead, lays out a few ideas for extracting metadata from user interaction with these new social multimedia applications, as the last step of the approach to social multimedia.

7 [Step 4] Leveraging user interaction

Yes, implicit relevance feedback has been with us for a while [7]. Nevertheless, implicit relevance feedback via user interaction offers a specific opportunity in social multimedia applications. In particular, three factors contribute to the opportunity: (1) narrower focus of the proposed applications that results in a reasonably predictable user intent; (2) richer interaction methods that can enable more refined feedback mechanisms; and (3) the possibility of feedback at the sub-resource level (e.g., parts of a video clip).

This section describes two possible methods for utilizing user interaction for implicit relevance feedback, in the Flickr Landmarks and Concert Sync systems. We had implemented and tested the Flickr Landmarks feedback ideas [4]. The

interaction and feedback described for Concert Sync are currently being implemented, and are described below to illustrate the generality of the fourth step.

7.1 [Flickr Landmarks] What's really representative?

The Flickr Landmarks application selects representative photos for landmarks. The landmark photos can be used for both search and visualization, and implicit feedback from both these environment can help refine and improve the set of displayed photos.

We implemented this idea in our World Explorer application [5] and tested it in later work [4]. The scenario tested in [4] was a visualization, where a user is shown the World Explorer map with overlaid tags that represent important features in that location (e.g., landmarks). When hovering over a tag with the mouse pointer, the application loaded 20 public Flickr photos that were annotated with that tag, from the geographic region where the tag appears; i.e., photos that visually explain and extend the tag information. In Fig. 1, the user hovered over the `Golden Gate Bridge` tag to see related photos. The photos are laid out in random ordering that provides an aesthetically pleasing view while intentionally obscuring some of the images. Once the photos are displayed, any photo can be expanded and examined in more detail by double-clicking on it. When expanded, the image is shown in correct rotation, together with additional metadata such as the photo title and the name of the user who took that photo. Users can also "close" a photograph by clicking on the "X" icon on the top right corner of the photo. Thus, we expect the user actions to provide feedback regarding relevancy: users are likely to expand the view of relevant, representative photos, while clicking to get irrelevant photos out of their way.

The data collection for this experiment had demonstrated the potential of social multimedia content filtering using implicit feedback. We logged the user interaction with this World Explorer application for 21 days. The numbers, based on the actions of over 2,400 users, initially suggest that users are more likely to examine images that are representative of the tag in question, and will usually ignore portrait images and other photos that are not representative of the tag. Also, the numbers indicate that users are more inclined to close portrait and non representative images than they are to examine them in detail.[5] The complete results are provided in [4].

7.2 [Concert Sync] Collaborative directors

The opportunity for implicit feedback based on user interaction with this content is even more exciting for the Concert Sync application, as the presentation includes streaming video and audio content. Let us examine the proposed interface and interaction shown in Fig. 2. As explained above, the figure shows multiple concurrent videos being played in synchronized fashion. At any time, the viewers can select any of the videos displayed on the left to be displayed larger, in the main frame on the right. The audio track could switch to play the audio of the selected clip, or play the best-quality audio for each segment, detected as described above (Section 6.2).

[5] Bewilderingly, some of the most examined photos included women in minimal clothing, even when the photos were not necessarily relevant to the location or the tag – proving that human factors are not always as predictable as researchers would hope. Or maybe they are.

This type of interaction lends itself immediately to refinement of content. The system could record the videos selected for display by different users at any given time. If there are indeed significant differences between the overlapping clips, at any given point more users will choose to view certain clips that are perhaps more interesting, or otherwise capture some relevant visual content. In aggregate, over time, enough user interaction could thus implicitly inform the playback of the video: at any given moment, the most "popular" video clip would be the focus of the playback.

On top of finding the best video or audio for each segment of the concert, other interaction data such as seek or "scrub" of video [79, 80] can further refine our metadata regarding interesting moments in the show. Of course, we can also create a new, explicit editing interface for the users to create new videos based on the existing content. As Shaw and Schmitz had shown [83], such "remix" environment could produce compelling content as well as provide clues about the original content.

The implicit knowledge that is captured by the interaction in this application could result, in essence, in a user-driven directing (or editing) of the content. Such collaborative, implicit curation will lead to better representation of raw social multimedia content. Once again, it should be noted that we had not implemented or tested the features mentioned above in Concert Sync; they are described here for completeness and to illustrate the generality of the approach.

8 Evaluation of social multimedia search and mining systems

This article discussed an approach for creating new social multimedia search and mining systems, but what is the proper way to evaluate such emerging systems? In our research group, we have grappled with various evaluation techniques and philosophies. In the root of the evaluation issue is that fact that when creating new experiences and prototypes, one cannot always boil the evaluation down to metrics [70]. For example, the user experience in our applications could be as much a factor for the "success" of the application as relevance, or precision and accuracy. The evaluation of the applications in context of specific user tasks is important, but the results are not always measurable, as the tasks in these new social multimedia applications are often not tied to clear performance metrics.

Some commonly used evaluation approaches may be lacking. Many research efforts, for example, have opted for a questionnaire-based approach, usually administering questionnaires to a small number of users of the system, often in comparison to another system the participant was exposed to in a within-subject design. These questionnaires are likely to include Leikert-scale questions about properties and qualities of the system (e.g., "enjoyable", "easy to use", "satisfying", "confusing", "likely to use it again") as well as open-ended comments. However, such responses are subject to bias and issues of reliability. Further, it is often the case that comparing these qualities across systems is meaningless; in many cases the new system is not "better" in any quantifiable way from another existing system. For example, comparing the Concert Sync browsing interface versus another, say, which plays the clips in succession, is akin to comparing a box of apples to a row of oranges.[6]

[6]It must be said that this author is also "guilty" in administrating such evaluations in past research.

Indeed, many of the required insights about such new experiences (as opposed to new algorithms) are hard to generate using quantitative methods: what are the different factors that are in play when users are interacting with the new application? How do the users perceive and understand the presentation and interaction mode? How do users feel the new experience helps (or hinders) them in performing their task? Some research efforts have used open-ended questions and questionnaires about a system in their evaluation; these are indeed an improvement over quantitative task measurements or specified questionnaire items, but this type of inquiry might still prove insufficient and is not likely to generate reproducible and generalizeable results.

We employed a two-pronged approach that can help in producing insights that will benefit other researchers and inform developers building similar or relevant applications. Short of employing (quantitative) large-scale analysis and (qualitative) rigorous observations called for by Shneiderman [84], we selected more modest evaluation goals that are more appropriate for emerging applications or new prototypes and systems. First, we performed a direct quantitative evaluation of the important system components. Second, we engaged in deep, extensive qualitative evaluation of the user's interaction with the system or interactive prototype.

The direct evaluation portion of our evaluation effort is based on the idea that output of specific computational portions of the system can be evaluated quantitatively. We established the different aspects of the system that could be measured quantitatively (and are significant enough for us to have cared). These components included, to name one example from each application, the selection of representative photos (Flickr Landmarks), and the evaluation of best-quality audio selection (Concert Sync). Note that in both these cases, as well as others, some human interpretation is needed for the evaluation: what is good audio quality? What is a representative photo? We used various data collection methods with human judges, mostly using answer forms submitted by multiple responders, to answer these questions in a robust and reliable way [45, 46] without exposing the respondents to the actual system.

For instance, another component of Concert Sync we evaluated in [46] is identifying important moments in the show, as explained above (Section 6.2). To evaluate the success in identifying those moments, we do not need to evaluate an interface or interaction, but instead evaluate the algorithm output. In that evaluation, we focused only on the clips that capture a clearly identifiable song in each concert (the song was manually identified). We compared our algorithm's ranking of each clip to the popularity of the respective track on Last.fm, a popular social music website with play chart and other track-level popularity data. We found a statistically significant correlation between these two values ($r^2 \sim .44$, $p < .001$, $N = 41$).

The second thrust of evaluation methods we deployed was a (mostly) qualitative analysis of the interactive systems we built. Creswell talks about Qualitative research as "means for exploring and understanding the meaning individuals or groups ascribe to a social or human problem" that "involves emerging questions and procedures, data analysis inductively building from particulars to general themes, and the researcher making interpretations of the meaning of the data." [26]. Indeed, we used qualitative methods to collect and analyze information from participants about our interactive social multimedia prototypes. To use Creswell's terms, our "human problem" is often defined in terms of the system's proposed goals, and it must be tied to user's goals and motivated and driven by existing user needs. The

"procedures" are a set of tasks that the users are likely to execute when using the system, and can be either simulated in lab settings or recreated from the interaction logs of existing users of such system.

We used the participant's interaction (guided by specific tasks, if performed in lab settings) to inform semi-structure interviews that allows the participant to talk about their actions and expectation. In lab settings, we can follow a "think-aloud" procedure where the participants describe the steps they are taking. Otherwise, using interaction logs, we could visualize the past activity of participants in detail that allows them to reflect on their actions (e.g., [58]). In both cases, the users can comment on their (past or present) interaction with the system such that the comments are grounded in the user's actual activity. We recorded the conversations with the participants, then analyzed them using Grounded Theory to identify emerging and recurrent themes. With this method, we could get significant, meaningful feedback about the system and the interaction, without limiting our analysis to pre-conceived measurements and questionnaire items.

For example, in evaluating World Explorer [5], we had invited 10–20 participants to interact with our system. We had identified in advance some real-world scenarios where a user is likely to use the system. We had the participants simulate these scenarios (e.g., "explore Paris in preparation to a future visit") and asked them to express their thoughts both during and after each part of the session. This strategy proved very useful in generating participant insights and feedback that we would not otherwise anticipate. That kind of feedback could greatly inform designers of similar systems and perhaps lead to more deductive, quantitative evaluations. In one example of our findings, based on the analysis of the aggregate comments from participants, we identified the need for "needle" mode to augment the "haystack"-like features of our visualization. In other words, our participants pointed out that the experience of visualizing the content was not always sufficient; when they were looking for specific items of interest, they needed to go deeper then the default items selected by the system. An evaluation in which the participants are instructed to execute a procedure that we know is afforded by the system and rate their experience would not have surfaced this requirement. The details of the procedure and results of our qualitative evaluation can be found in [5]. We also executed a similar evaluation on a mobile multimedia application [58] ; in other work (e.g. [3, 6]) we performed a more targeted evaluation, geared toward specific research questions about our social multimedia prototypes (e.g., the purpose of tagging and approach to privacy), but based on similar principals.

9 Conclusions

We live in an exciting time for multimedia research, as the "age of social multimedia" ushers in rapid changes in the amount and type of available content, in the features and depth of the metadata, in the platforms that run multimedia applications, as well as in the applications themselves. These changes call for new challenges that can leverage the new trends, perhaps in addition to using a renewed opportunity to iterate on existing multimedia problems.

This article demonstrated one approach that proved successful in this domain of social multimedia. Taking this approach allowed us to build scalable real-world

applications that leverage multimedia content analysis in a robust manner, touching on two ideas from widely divergent application domains, video concert videos and landmark images. What other applications could be created using the same approach? Capable researchers of the multimedia community will certainly be able to provide various answers to this question. Several possible areas, which are at the same time interesting and ripe for new applications, are: citizen participation and local government (where the location context is significant); photo-driven environmental sensing applications; hyper-local interaction spaces such as museums; social multimedia-based memory or collection augmentation; disaster aid and documentation and so forth.

Indeed, future work in this area will need to take a human-centered approach to designing and developing the new multimedia applications. We need additional ethnographic and exploratory work to understand people's actions and intentions in various existing and new settings. In addition, we need to continue work on more efficient and scalable algorithms that will allow an order of magnitude improvement for both context and content analysis. Such systems will allow better adaptation and personalization of social multimedia content.

Acknowledgements The author would like to acknowledge the contributions of his colleagues and interns at Yahoo! Research Berkeley, whose ideas, expertise and excitement made this work possible—or, in fact, made this work [period]. In particular, Lyndon Kennedy made many of the key contributions described here.

References

1. Abbasi R, Chernov S, Nejdl W, Paiu R, Staab S (2009) Exploiting flickr tags and groups for finding landmark photos. In: ECIR '09: proceedings of the 31th European conference on ir research on advances in information retrieval. Springer-Verlag, Berlin, Heidelberg, pp 654–661
2. Adams B, Phung D, Venkatesh S (2006) Extraction of social context and application to personal multimedia exploration. In: MULTIMEDIA '06: proceedings of the 14th annual ACM international conference on multimedia. ACM, New York, NY, USA, pp 987–996
3. Ahern S, Eckles D, Good N, King S, Naaman M, Nair R (2007) Over-exposed? privacy patterns and considerations in online and mobile photo sharing. In: CHI '07: proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, NY, USA
4. Ahern S, King S, Naaman M, Nair R (2007) Summarization of online image collections via implicit feedback. In WWW '07: proceedings of the 16th international conference on World Wide Web. ACM, New York, NY, USA, pp 1325–1326
5. Ahern S, Naaman M, Nair R, Yang JH-I (2007) World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In JCDL '07: proceedings of the seventh ACM/IEEE-CS joint conference on digital libraries. ACM, New York, NY, USA, pp 1–10
6. Ames M, Naaman M (2007) Why we tag: motivations for annotation in mobile and online media. In: CHI '07: proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, NY, USA
7. Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. Addison Wesley
8. Becker H, Naaman M, Gravano L (2009) Event identification in social media. In: WebDB '09: proceedings of the 12th international workshop on the web and databases: colocated with ACM SIGMOD
9. Becker H, Naaman M, Gravano L (2010) Learning similarity metrics for event identification in social media. In: WSDM'10: proceedings of the third ACM international conference on web search and data mining. ACM, New York, NY, USA, pp 291–300
10. Beerends JG, Stemerdink JA (1992) A perceptual audio quality measure based on a psychoacoustic sound representation. J Audio Eng Soc 40(12):963–978
11. Berg TL, Forsyth DA (2007) Automatic ranking of iconic images. Technical report, U.C. Berkeley

12. Beyer KS, Goldstein J, Ramakrishnan R, Shaft U (1999) When is "nearest neighbor" meaningful? In ICDT '99: proceedings of the 7th international conference on database theory. Springer-Verlag, London, UK, pp 217–235
13. Boll S (2007) Multitube–where web 2.0 and multimedia could meet. IEEE Multimed 14(1):9–13
14. Boll S, Bulterman D, Jain R, Chua T-S, Lienhart R, Wilcox L, Davis M, Venkatesh S (2004) Between context-aware media capture and multimedia content analysis: where do we find the promised land? In MULTIMEDIA '04: proceedings of the 12th annual ACM international conference on multimedia. ACM, New York, NY, USA, pp 868–868
15. Boll S, Sandhaus P, Scherp A, Thieme S (2006) Metaxa—context- and content-driven metadata enhancement for personal photo books. In: Advances in multimedia modeling, pp 332–343
16. Boll S, Sandhaus P, Scherp A, Westermann U (2007) Semantics, content, and structure of many for the creation of personal photo albums. In: MULTIMEDIA '07: proceedings of the 15th international conference on multimedia. ACM, New York, NY, USA, pp 641–650
17. Boutemedjet S, Ziou D (2008) A graphical model for context-aware visual content recommendation. IEEE Trans Multimedia 10(1):52–62
18. Bulterman DCA (2004) Is it time for a moratorium on metadata? IEEE MultiMed 11(4):10–17
19. Cao L, Luo J, Huang TS (2008) Annotating photo collections by label propagation according to multiple similarity cues. In: MM '08: proceeding of the 16th ACM international conference on multimedia. ACM, New York, NY, USA, pp 121–130
20. Chang E (2008) Organizing multimedia data socially. In: CIVR '08: proceedings of the 2008 international conference on content-based image and video retrieval. ACM, New York, NY, USA, pp 569–570
21. Chang EY (2005) Extent: fusing context, content, and semantic ontology for photo annotation. In: CVDB '05: proceedings of the 2nd international workshop on computer vision meets databases. ACM, New York, NY, USA, pp 5–11
22. Chen W-C, Battestini A, Gelfand N, Setlur V (2009) Visual summaries of popular landmarks from community photo collections. In: MM '09: proceedings of the seventeen ACM international conference on multimedia. ACM, New York, NY, USA, pp 789–792
23. Choudhury MD, Sundaram H, John A, Seligmann DD (2009) What makes conversations interesting? Themes, participants and consequences of conversations in online social media. In: WWW '09: proceeding of the 18th international conference on World Wide Web. ACM, New York, NY, USA
24. Christel MG, Hauptmann AG, Wactlar HD (2002) Collages as dynamic summaries for news video. In: MULTIMEDIA '02: proceedings of the 10th international conference on multimedia. ACM, pp 561–569
25. Crandall D, Backstrom L, Huttenlocher D, Kleinberg J (2009) Mapping the world's photos. In: WWW '09: proceeding of the 18th international conference on World Wide Web. ACM, New York, NY, USA
26. Creswell JW (2002) Research design: qualitative, quantitative, and mixed methods approaches, 2nd edn. Sage Publications, Thousand Oaks, CA, USA
27. Cunningham SJ, Nichols DM (2008) How people find videos. In: JCDL '08: proceedings of the Eigth ACM/IEEE joint conference on digital libraries. ACM, New York, NY, USA
28. Das M, Farmer J, Gallagher A, Loui A (2008) Event-based location matching for consumer image collections. In: CIVR '08: proceedings of the 2008 international conference on content-based image and video retrieval. ACM, New York, NY, USA, pp 339–348
29. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):1–60
30. Davis M, King S, Good N, Sarvas R (2004) From context to content: leveraging context to infer media metadata. In: Proceedings of the 12th international conference on multimedia (MM2004). ACM, pp 188–195
31. Davis M, Smith M, Stentiford F, Bambidele A, Canny J, Good N, King S, Janakiraman R (2006) Using context and similarity for face and location identification. In: Proceedings of the IS&T/SPIE 18th annual symposium on electronic imaging science and technology
32. Dubinko M, Kumar R, Magnani J, Novak J, Raghavan P, Tomkins A (2006) Visualizing tags over time. In: WWW '06: proceedings of the 15th international conference on World Wide Web. ACM, New York, NY, USA, pp 193–202
33. Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley-Interscience
34. Elliott B, Özsoyoglu ZM (2008) Annotation suggestion and search for personal multimedia objects on the web. In: CIVR '08: proceedings of the 2008 international conference on content-based image and video retrieval. ACM, New York, NY, USA, pp 75–84

35. Flickr.com (2010) http://www.flickr.com
36. Graham A, Garcia-Molina H, Paepcke A, Winograd T (2002) Time as essence for photo browsing through personal digital libraries. In: JCDL '02: proceedings of the second ACM/IEEE-CS joint conference on digital libraries
37. Haitsma J, Kalker T (2003) A highly robust audio fingerprinting system with an efficient search strategy. J New Music Res 32(2):211–221
38. Hao Q, Cai R, Wang XJ, Yang JM, Pang Y, Zhang L (2009) Generating location overviews with images and tags by mining user-generated travelogues. In: MM '09: proceedings of the seventeen ACM international conference on multimedia. ACM, New York, NY, USA, pp 801–804
39. Jaffe A, Naaman M, Tassa T, Davis M (2006) Generating summaries and visualization for large collections of geo-referenced photographs. In: MIR '06: proceedings of the 8th ACM international workshop on multimedia information retrieval. ACM, New York, NY, USA, pp 89–98
40. Jaimes A, Christel M, Gilles S, Sarukkai R, Ma W-Y (2005) Multimedia information retrieval: what is it, and why isn't anyone using it? In: MIR '05: proceedings of the 7th ACM SIGMM international workshop on multimedia information retrieval. ACM, New York, NY, USA, pp 3–8
41. Ji R, Xie X, Yao H, Ma W-Y (2009) Mining city landmarks from blogs by graph modeling. In: MM '09: proceedings of the seventeen ACM international conference on multimedia. ACM, New York, NY, USA, pp 105–114
42. Jing F, Zhang L, Ma W-Y (2006) Virtualtour: an online travel assistant based on high quality images. In: Proceedings of the 14th international conference on multimedia (MM2005). ACM, New York, NY, USA, pp 599–602
43. Joshi D, Luo J (2008) Inferring generic activities and events from image content and bags of geotags. In: Proceedings of the 2008 international conference on content-based image and video retrieval. ACM, Niagara Falls, Canada, pp 37–46
44. Kennedy LS, Chang S-F (2007) A reranking approach for context-based concept fusion in video indexing and retrieval. In: CIVR '07: proceedings of the 6th ACM international conference on image and video retrieval. ACM, New York, NY, USA, pp 333–340
45. Kennedy LS, Naaman M (2008) Generating diverse and representative image search results for landmarks. In: WWW '08: proceeding of the 17th international conference on World Wide Web. ACM, New York, NY, USA, pp 297–306
46. Kennedy L, Naaman M (2009) Less talk, more rock: automated organization of community-contributed collections of concert videos. In: WWW '09: proceeding of the 18th international conference on World Wide Web. ACM, New York, NY, USA
47. Kennedy L, Chang S-F, Kozintsev I (2006) To search or to label? Predicting the performance of search-based automatic image classifiers. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval, pp 249–258
48. Kennedy L, Naaman M, Ahern S, Nair R, Rattenbury T (2007) How flickr helps us make sense of the world: context and content in community-contributed media collections. In: Proceedings of the 15th international conference on multimedia (MM2007). ACM, New York, NY, USA, pp 631–640
49. Lew MS, Sebe N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: state of the art and challenges. ACM TOMCCAP 2(1):1–19
50. Liu Y, Zhang D, Lu G, Ma W-Y (2009) A survey of content-based image retrieval with high-level semantics. Pattern Recogn. 40(1):262–282
51. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
52. Luo J, Boutell M, Brown C (2006) Pictures are not taken in a vacuum—an overview of exploiting context for semantic scene content understanding. IEEE Signal Process Mag 23(2):101–114
53. Manjunath BS, Ma WY (1996) Texture features for browsing and retrieval of image data. IEEE Trans Pattern Anal Mach Intell 18(8):837–842
54. Mertens R, Farzan R, Brusilovsky P (2006) Social navigation in web lectures. In: HYPERTEXT '06: proceedings of the seventeenth conference on hypertext and hypermedia. ACM, New York, NY, USA, pp 41–44
55. Naaman M, Nair R (2008) ZoneTag's collaborative tag suggestions: what is this person doing in my phone? IEEE Multimed 15(3):34–40
56. Naaman M, Garcia-Molina H, Paepcke A, Yeh RB (2005) Leveraging context to resolve identity in photo albums. In: JCDL '05: proceedings of the Fifth ACM/IEEE-CS joint conference on digital libraries. ACM Press

57. Naaman M, Harada S, Wang Q, Garcia-Molina H, Paepcke A (2004) Context data in geo-referenced digital photo collections. In: Proceedings of the 12th international conference on multimedia (MM2004). ACM
58. Naaman M, Nair R, Kaplun V (2008) Photos on the go: a mobile application case study. In: CHI '08: proceeding of the twenty-sixth annual SIGCHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 1739–1748
59. Naaman M, Paepcke A, Garcia-Molina H (2003) From where to what: metadata sharing for digital photographs with geographic coordinates. In: 10th international conference on cooperative information systems (CoopIS)
60. Naaman M, Song YJ, Paepcke A, Garcia-Molina H (2004) Automatic organization for digital photographs with geographic coordinates. In:JCDL '04: proceedings of the fourth ACM/IEEE-CS joint conference on digital libraries
61. Naci SU, Hanjalic A (2007) Intelligent browsing of concert videos. In: MULTIMEDIA '07: proceedings of the 15th international conference on multimedia. ACM, New York, NY, USA, pp 150–151
62. Nair R, Reid N, Davis M (2005) Photo LOI: browsing multi-user photo collections. In: Proceedings of the 13th international conference on multimedia (MM2005). ACM
63. Negoescu RA, Gatica-Perez D (2008) Analyzing flickr groups. In: CIVR '08: proceedings of the 2008 international conference on content-based image and video retrieval. ACM, New York, NY, USA, pp 417–426
64. Negoescu R-A, Adams B, Phung D, Venkatesh S, Gatica-Perez D (2009) Flickr hypergroups. In: MM '09: proceedings of the seventeen ACM international conference on multimedia. ACM, New York, NY, USA, pp 813–816
65. Nov O, Naaman M, Ye C (2008) What drives content tagging: the case of photos on flickr. In: CHI '08: proceeding of the twenty-sixth annual SIGCHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 1097–1100
66. O'Hare N, Smeaton AF (2009) Context-aware person identification in personal photo collections. IEEE Trans Multimedia 11(2):220–228
67. O'Hare N, Gurrin C, Jones GJF, Smeaton AF (2005) Combination of content analysis and context features for digital photograph retrieval. In: 2nd IEE European workshop on the integration of knowledge, semantic and digital media technologies
68. O'Hare N, Gurrin C, Lee H, Murphy N, Smeaton AF, Jones GJF (2005) My digital photos: Where and when? In: Proceedings of the 13th international conference on multimedia (MM2005). ACM
69. O'Hare N, Lee H, Cooray S, Gurrin C, Jones G, Malobabic J, O'Connor N, Smeaton A, Uscilowski B (2006) MediAssist: using content-based analysis and context to manage personal photo collections. In: Image and video retrieval, pp 529–532
70. Olsen DR Jr (2007) Evaluating user interface systems research. In: UIST '07: proceedings of the 20th annual ACM symposium on user interface software and technology, pp 251–258
71. Paillard B, Mabilleau P, Morissette S, Soumagne J (1992) PERCEVAL: perceptual evaluation of the quality of audio signals. J Audio Eng Soc 40(1/2):21–31
72. Palmer S, Rosch E, Chase P (1981) Canonical perspective and the perception of objects. In: Long JB, Baddeley AD (eds) Attention and performance IX. Lawrence Erlbaum Associates, Hillsdale, N.J., pp 135–151
73. Pigeau A, Gelgon M (2004) Organizing a personal image collection with statistical model-based ICL clustering on spatio-temporal camera phone meta-data. J Vis Commun Image Represent 15(3):425–445
74. Qi G-J, Hua X-S, Zhang H-J (2009) Learning semantic distance from community-tagged media collection. In: MM '09: proceedings of the seventeen ACM international conference on multimedia. ACM, New York, NY, USA, pp 243–252
75. Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of the thirtieth annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, pp 103–110
76. Salovaara A, Jacucci G, Oulasvirta Timo Saari A, Kanerva P, Kurvinen E, Tiitta S (2006) Collective creation and sense-making of mobile media. In: CHI '06: proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 1211–1220
77. Schmitz P (2006) Inducing ontology from flickr tags. In: Proceedings of the workshop on collaborative web tagging at WWW2006

78. Setz AT, Snoek CGM (2009) Can social tagged images aid concept-based video search? In: ICME'09: proceedings of the 2009 IEEE international conference on multimedia and Expo. IEEE Press, Piscataway, NJ, USA, pp 1460–1463
79. Shamma DA, Shaw R, Shafton PL, Liu Y (2007) Watch what I watch: using community activity to understand content. In: MIR '07: proceedings of the international workshop on workshop on multimedia information retrieval. ACM, New York, NY, USA, pp 275–284
80. Shamma DA, Bastea-Forte M, Joubert N, Liu Y (2008) Enhancing online personal connections through the synchronized sharing of online video. In: CHI '08: CHI '08 extended abstracts on human factors in computing systems. ACM, New York, NY, USA, pp 2931–2936
81. Shamma DA, Kennedy L, Churchill EF (2009) Tweet the debates: understanding community annotation of uncollected sources. In: WSM '09: proceedings of the first SIGMM workshop on Social media. ACM, New York, NY, USA, pp 3–10
82. Shamma DA, Kennedy L, Churchill E (2010) Statler: summarizing media through short-message services. In: CSCW '10: proceedings of the 2010 ACM conference on computer supported cooperative work. ACM, New York, NY, USA
83. Shaw R, Schmitz P (2006) Community annotation and remix: a research platform and pilot deployment. In: HCM '06: proceedings of the 1st ACM international workshop on human-centered multimedia. ACM, New York, NY, USA, pp 89–98
84. Shneiderman B (2008) COMPUTER SCIENCE: Science 2.0. Science 319(5868):1349–1350
85. Shrestha P, Weda H, Barbieri M, Sekulovski D (2006) Synchronization of multiple video recordings based on still camera flashes. In: MULTIMEDIA '06: proceedings of the 14th international conference on multimedia. ACM, pp 137–140
86. Shrestha P, Barbieri M, Weda H (2007) Synchronization of multi-camera video recordings based on audio. In: MULTIMEDIA '07: proceedings of the 15th international conference on multimedia. ACM, pp 545–548
87. Sigurbjörnsson B, van Zwol R (2008) Flickr tag recommendation based on collective knowledge. In: WWW '08: proceeding of the 17th international conference on World Wide Web. ACM, New York, NY, USA, pp 327–336
88. Simon I, Snavely N, Seitz SM (2007) Scene summarization for online image collections. In: ICCV '07: proceedings of the 11th IEEE international conference on computer vision.
89. Sinha P, Jain R (2008) Classification and annotation of digital photos using optical context data. In: CIVR '08: proceedings of the 2008 international conference on content-based image and video retrieval. ACM, New York, NY, USA, pp 309–318
90. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
91. Snavely N, Garg R, Seitz SM, Szeliski R (2008) Finding paths through the world's photos. In: SIGGRAPH '08: proceedings of ACM SIGGRAPH. ACM, New York, NY, USA, pp 1–11
92. Snoek CGM, Worring M, Smeulders AWM, Freiburg B (2007) The role of visual content and style for concert video indexing. In: IEEE international conference on multimedia and expo, pp 252–255, 2–5 July 2007
93. Strickerand M, Orengo M (1995) Similarity of color images. In: Proc. SPIE storage and retrieval for image and video databases vol 2420, pp 381–392
94. Syeda-Mahmood T, Ponceleon D (2001) Learning video browsing behavior and its application in the generation of video previews. In: MULTIMEDIA '01: proceedings of the 9th ACM international conference on multimedia. ACM, New York, NY, USA, pp 119–128
95. Tang J, Yan S, Hong R, Qi G-J, Chua T-S (2009) Inferring semantic concepts from community-contributed images and noisy tags. In: MM '09: proceedings of the seventeen ACM international conference on multimedia. ACM, New York, NY, USA, pp 223–232
96. Thiede T, Treurniet WC, Bitto R, Schmidmer C, Sporer T, Beerends JG, Colomes C (2000) PEAQ-the ITU standard for objective measurement of perceived audio quality. J Audio Eng Soc 48(1/2):3–29
97. Toyama K, Logan R, Roseway A (2003) Geographic location tags on digital images. In: Proceedings of the 11th international conference on multimedia (MM2003). ACM, pp 156–166
98. Tsai C-M, Qamra A, Chang E (2005) Extent: inferring image metadata from context and content. In: IEEE International conference on multimedia and expo
99. Uchihashi S, Foote J, Girgensohn A (1999) Video manga: generating semantically meaningful video summaries. In: MULTIMEDIA '99: proceedings of the 7th international conference on multimedia. ACM, pp 383–392
100. van Houten Y, Naci U, Freiburg B, Eggermont R, Schuurman S, Hollander D, Reitsma J, Markslag M, Kniest J, Veenstra M, Hanjalic A (2005) The multimedian concert-video browser.

In: IEEE international conference on multimedia and expo, 2005. ICME, pp 1561–1564, 6-6 July 2005
101. Wang A (2003) An industrial strength audio search algorithm. In: Proceedings of the international conference on music information retrieval
102. Wang M, Yang K, Hua X-S, Zhang H-J (2009) Visual tag dictionary: interpreting tags with visual words. In: WSMC '09: proceedings of the 1st workshop on web-scale multimedia corpus. ACM, New York, NY, USA, pp 1–8
103. Westermann U, Jain R (2007) Toward a common event model for multimedia applications. IEEE Multimed 14(1):19–29
104. Wu Y, Chang EY, Tseng BL (2005) Multimodal metadata fusion using causal strength. In: Proceedings of the 13th international conference on multimedia (MM2005). ACM, New York, NY, USA, pp 872–881
105. Wu L, Hua X-S, Yu N, Ma W-Y, Li S (2008) Flickr distance. In: MM '08: proceeding of the 16th ACM international conference on multimedia. ACM, New York, NY, USA, pp 31–40
106. Wu L, Hoi SCH, Jin R, Zhu J, Yu N (2009) Distance metric learning from uncertain side information with application to automated photo tagging. In: MM '09: proceedings of the seventeen ACM international conference on multimedia. ACM, New York, NY, USA, pp 135–144
107. YouTube (2010) http://youtube.com/
108. Zheng Y-T, Zhao M, Song Y, Adam H, Buddemeier U, Bissacco A, Brucher F, Chua T-S, Neven H (2009) Tour the world: building a web-scale landmark recognition engine. In: CVPR '09: IEEE conference on computer vision and pattern recognition, pp 1085–1092
109. Zheng Y-T, Zhao M, Song Y, Adam H, Buddemeier U, Bissacco A, Brucher F, Chua T-S, Neven H (2009) Tour the world: building a web-scale landmark recognition engine. In: CVPR workshops 2009: IEEE computer society conference on computer vision and pattern recognition workshops, 2009, pp 1085–1092
110. Zunjarwad A, Sundaram H, Xie L (2007) Contextual wisdom: social relations and correlations for multimedia event annotation. In: Proceedings of the 15th international conference on multimedia, pp 615–624

Mor Naaman is an assistant professor at the Rutgers University School of Communication and Information. His research interests include social information systems, social media, multimedia and mobile computing. Prior to joining Rutgers, Mor worked as a research scientist at Yahoo! Research Berkeley, where he led a team of research engineers and interns investigating the future of mobile and social media technology. Mor received a PhD in Computer Science from Stanford University. His research in the Stanford Infolab also focused on digital media, and in particular the management of digital photographs, thereby allowing (and requiring!) him to take photos throughout his research career. Mor co-founded the ACM Multimedia Grand Challenge, and served as a co-chair of the JCDL 2008 Program Committee. He is a recipient of a Google Research Award, and three best paper awards. In previous careers, Mor was a professional basketball player as well as a software developer and a college radio DJ. In subsequent careers, Mor hopes to be a professional backpacker and traveler.