Less Talk, More Rock: Automated Organization of Community-Contributed Collections of Concert Videos

Lyndon Kennedy
Yahoo! Research
Santa Clara, CA
lyndonk@yahoo-inc.com

Mor Naaman *
Dept. of Library and Information Science
Rutgers University, New Brunswick, NJ
mor@scils.rutgers.edu

ABSTRACT

We describe a system for synchronization and organization of user-contributed content from live music events. We start with a set of short video clips taken at a single event by multiple contributors, who were using a varied set of capture devices. Using audio fingerprints, we synchronize these clips such that overlapping clips can be displayed simultaneously. Furthermore, we use the timing and link structure generated by the synchronization algorithm to improve the findability and representation of the event content, including identifying key moments of interest and descriptive text for important captured segments of the show. We also identify the preferred audio track when multiple clips overlap. We thus create a much improved representation of the event that builds on the automatic content match. Our work demonstrates important principles in the use of content analysis techniques for social media content on the Web, and applies those principles in the domain of live music capture.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Algorithms, Human Factors

Keywords

audio fingerprinting, video, social media, synchronization

1. INTRODUCTION

It happened just last night – you attended Iron Maiden's show at the Pavilion, part of the band's Somewhere Back in Time tour. Slowly getting over your hangover and wanting to re-live some of the best moments from last night, you make your way to the computer to see if people uploaded video clips captured at the show. Unfortunately, there are too many clips. You sift through sixteen different video captures of "Number of the Beast," interleaved with videos of people drinking beer before the show, before giving up on finding additional interesting moments.

The above scenario is a likely and perhaps frequent occurrence on the Web today [7]. The availability of video capture devices and the high reach and impact of social video sharing sites like YouTube [23] make video content from live shows relatively easy to share and find [2]. Users of YouTube share millions of videos capturing live musical performances, from classical pianist Ivo Pogorelich to metal rockers Iron Maiden. Potentially, such an abundance of content could enable a comprehensive and deeper multimedia coverage of captured events. However, there are new challenges that impede this new potential: the sample scenario above, for example, illustrates issues of relevance, findability, and redundancy of content.

The lack of detailed metadata associated with video content presents several interesting challenges. First, with no accurate, semantic event-based metadata, it is not trivial to automatically identify a set of video clips taken at a given event with high recall and precision. Second, with no dependable time-based metadata associated with the clips, aligning and synchronizing the video clips from the same event cannot be done using simple timestamps.

In this paper, we report on an approach for solving the synchronization problem, and how we leverage the synchronization data to extract additional metadata. The metadata would help us organize and present video clips from live music shows. We start by assuming the existence of a curated set of clips, having already identified (with reasonable precision) the video clips from each event (we do not report here on our system that automatically crawls the Web to retrieve those clips, but note that our processing methods here should be relatively robust to false positives).

Given all the video clips captured by users at a certain show, we use audio fingerprinting [4, 21] to synchronize the content. In other words, we use the clips' audio tracks to detect when the same moment is captured in two different videos, identify the overlap, and specify the time offset between any pair of overlapping clips. We note that while audio fingerprinting is not a new technique, we apply the technique here in novel ways.

The synchronization of clips allows us to create 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 1 presents one possible viewing interface.

Beyond synchronized playback, the synchronization and overlap data help improve both findability and relevance of clips from the event. Once synchronized, we use both the relative time information and links between overlapping

*This work was done while the second author was at Yahoo!.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

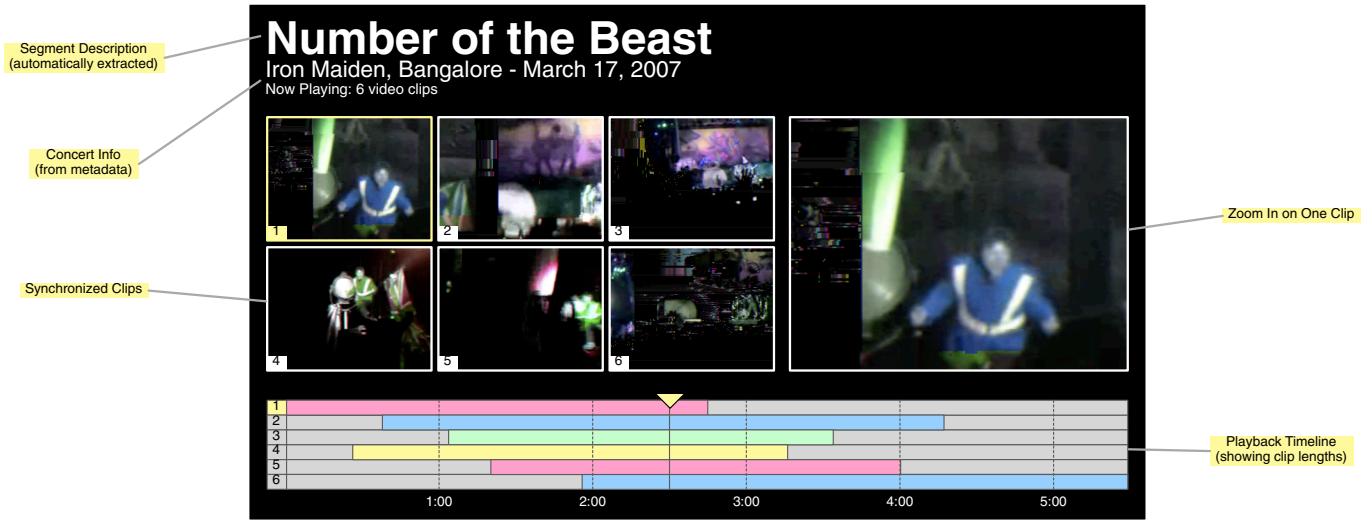

Figure 1: A sample interface for synchronized playback of concert video clips.

clips to generate important metadata about the clips and the event. First, we show how we identify level of interest [11] and significant moments in the show as captured by the users. Second, we mine the tags of videos associated with a single point in time to extract semantically meaningful descriptive terms for the key moments in the show; these terms can be used to represent or explain the aggregated content. Third, we use the link structure created by the audio fingerprinting when a clip matches another to find the highest-quality audio recording of any time segment, given multiple overlapping recordings.

To summarize, the specific contributions of this work are as follows:

- A framework for identifying and synchronizing video clips taken at the same event using audio fingerprinting.
- An initial analysis of the audio synchronization and its effectiveness under different conditions.
- An approach to leveraging the content-based match to extract meaningful metadata about the event, including time-based points of interest and descriptive text that can be used for retrieval or summarization.
- Demonstrating and evaluating how synchronization links between clips can be utilized to find the highest quality audio content.

Our primary focus in this work is an in-depth exploration of the different methods, rather than building and evaluating a browsing system. We examine the performance and effectiveness of our approach for using audio fingerprinting to synchronize and generate metadata about a set of videos from an event. The construction and human-centered evaluation of a browsing system utilizing these cues is outside the scope of this work, and will be pursued in future work.

We describe the application of audio fingerprinting to the domain of live music videos in Section 3. In Section 4 we report on the new techniques for the generation of metadata based on the synchronization data. An evaluation, using a real dataset, of both the synchronization and the new applications of the synchronization data is provided in Section 5. We begin by reporting on important related work.

2. RELATED WORK

We refer to previous work in a number of related areas, including: event-based management of media, research on video summarization, and work on multimedia related to live music or concerts. In addition, we report on audio fingerprinting research and some existing applications.

Our work contributes to a significant body of work in event-based management of media. Most of this work considered personal events in stand-alone, personal systems (e.g., [3, 9] and more). Lately, the event construct was expanded to include social web-based representation [24] and other aspects of event modeling [22].

Projects related to video summarization and selection of keyframes were mostly based on content analysis [1, 19], but some community-based methods have recently been proposed. For instance, Shaw and Schmitz [15] suggest that community "remix" data could help in similar summarization tasks. The authors show how knowledge can be extracted from the patterns that emerged in remix activity of individual clips. Our approach in this work is indeed related to Shaw and Schmitz, also leveraging community activity (i.e., recording) to learn about a video collection. A different approach was proposed by [14], that suggested using viewing activity to reason about content. Our model and application scenario are, of course, widely divergent from all of the above, yet can potentially be used for similar tasks.

A number of research efforts have addressed the domain of media from live music events [10, 18, 20]. 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). In [10], Naci and Hanjalic provide a demo that utilizes audio analysis to help users find interesting moments from the concert. Detecting interesting moments automatically is approximated by detecting applause, instrument solos and audio level of excitement; a browsing interface is provided that is supported by the extracted data. In [18], the authors use the visual signal of produced content from live concerts to create concept detectors including 'audience', 'drummer', 'stage' and so forth. In this work, we are not using the rarely-available produced

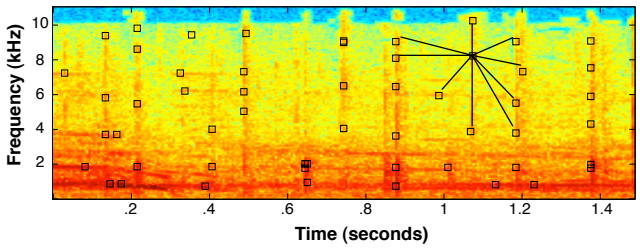

Figure 2: Spectrogram of a segment of audio, with selected landmarks overlaid (as black squares). One landmark demonstrates the fan-out to 9 adjacent landmarks (black lines).

content, but instead focus on community-contributed content widely available on the Web. However, a combination of produced and amateur sources may prove useful, as we note in the Future Work Section.

Most closely related to our work here, Shrestha et al. made significant contributions in synchronizing and aligning multiple video clips of a single scene. The authors first aligned video clips from multiple contributors using the detection of camera flashes [17]. Later, Shrestha et al. used audio fingerprints [13] (much like we have in this work) to accomplish a similar task. Our work differs not in the underlying technology, but in the application of the technology to extracting additional information. We shift the focus of our system from developing matching algorithms, like in [13] and [21], and focus on mining the structure of the discovered overlaps and audio re-use to create compelling new ways of aggregating and organizing community-contributed Web data.

Finally, this work builds on a core audio identification technology known as audio fingerprinting [4, 21], a rather robust and powerful technology that is already a reality in several commercial applications. In audio fingerprinting, audio recordings are characterized by local occurrences of particular structures. Given two recordings, one could rapidly identify if they were derived from the same original source material, despite a rather large amount of additive noise. In this work, we do not re-invent audio fingerprinting. Instead, we implement existing fingerprinting methods and reimagine them in an application not driven towards example-based search (as is typically done), but rather on mining large collections to find further aspects of real-world events which are captured and shared by many participants.

Several research efforts build on audio fingerprinting techniques. For example, in [5, 6], the author applies audio fingerprints as a sparse representation of pieces of audio already encountered in audio streams. He then mines this data to discover and segment "repeated events," which might be individual songs on a radio station, or commercials on a television station. In [12], the authors use audio fingerprints for identifying repeated events in personal audio recordings of experiences encountered by an individual throughout a day. Our work differs in that we do not identify *repeated* events, but actually *identical* events captured by multiple devices. Rather than monitoring a single stream of audio data, we monitor a data source created by dozens of authors, and apply the results in a novel application.

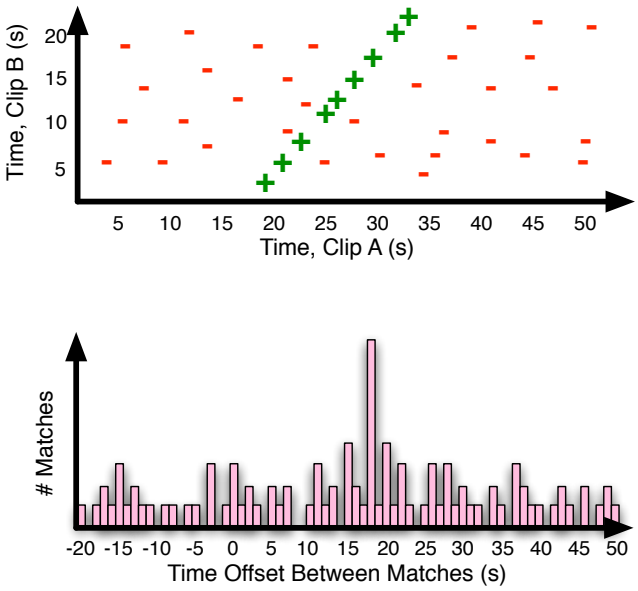

Figure 3: Visualization of matching between two audio clips. The presence of many matches along a diagonal line (top) indicates a high likelihood of a match between clips. The e ect can be better observed by taking the histogram of match frequency over various time di erence between matches (bottom).

3. ALIGNING VIDEO CLIPS WITH AUDIO FINGERPRINTS

While the audio fingerprinting technique is not one of our contributions here, we describe the method in brief for completeness. We outline the specific parameterization and settings we have used to apply the techniques to matching audio clips of the same event from multiple contributors. Some specific challenges in this application domain are the need to adaptively set detection thresholds that are resilient to multiple noisy sources, which we discuss in Section 3.2. Also, we propose an approach to *synchronizing* clips based on the matching detection results, which was not reported elsewhere as of yet.

3.1 Generating Fingerprints

To compute audio fingerprints for audio segments, we apply the method proposed by Wang [21]. The approach operates by taking the short-time Fourier 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 landmarks in an adjacent target zone. A visualization of the fingerprinting process is shown in Figure 2.

The power of this approach lies in its robustness against noise. The landmarks are in areas of high energy in the spectrogram, and are therefore the regions least likely to degrade in the presence of noise. And even if some landmarks are lost in the presence of noise, the matching algorithm (which

we explain in the following section) does not require accurate detection of all landmarks and hashes, and is, in fact, robust against the loss of the vast majority of landmarks.

In this work, the frequency and time values are each quantized into 64 values, or 6 bits. Each fingerprint consists of a concatenation of the frequencies of the two landmarks and the time difference between them, yielding an 18-bit hash, with 262,144 possible values. Landmarks are extracted over 1-second audio clips and fingerprints are calculated using a fixed density of 18 anchor points and a fan-out to 9 adjacent landmarks, giving 162 fingerprints per second of audio.

3.2 Matching Clips

The result of the above-described 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 done by finding all occurrences of matching hash values between the two clips. Since there is a fixed (and somewhat small) vocabulary for hash values, it is typical, and highly likely, that many spurious matches will be found between many points in the two clips. Intuitively, however, since audio recordings are constrained to progress at the same speed, 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.

This phenomenon can be observed in Figure 3. At the top, we see scatter plots of the times at which matching hash values are found across two clips, represented by 'plus' and 'minus' signs. We can see that a clear majority of the matches occur along a diagonal line with slope equal to 1 (those matches are marked with a 'plus' for visibility – symbolizing true positives). Viewing a histogram of the offsets between matching hash values between two clips (as shown in the bottom of Figure 3), can be more revealing: a very strong peak is seen in the offset histogram, suggesting a match between the two clips at the given offset.

In our work, we group the offsets of the found hash matches into 50ms bins (a size chosen to be shorter than a single frame in web videos, which tend to have frame rates lower than 15 frames per second) and normalize the counts in the bins by the sample mean and standard deviation of the counts of all bins. We compute a threshold based on the computed mean and standard deviation. We then determine whether or not the two clips are a match by checking whether any of the bins exceeds the threshold (which we vary in the evaluation).

Successful detection of overlapping clips relies on a few important factors. First, the less noise there is in a clip (or in both), the more likely it is that detection will be successful. The length of the clip also plays a role, where longer clips have more of a chance of generating correct matches. In our experience, very clean recordings can be matched with only about 10 seconds of audio. Moderately noisy clips would require closer to 30 seconds. Very noisy clips might require several minutes, or may never succeed.

3.3 Synchronization

Audio fingerprinting applications are typically focused on the detection of matches between two clips, either for the purposes of retrieval or the detection of duplicates. However, the detection process can also yield a sufficiently fine-grained estimation of the actual offset between two clips. Specifically, if a peak is detected by the matching algorithm, the value of the offset between the clips at that peak provides the offset needed to synchronize the two clips. Setting the bins' span to 50 ms means that when handling Web video clips (which typically have a low frame rate) this resolution is sufficient to synchronize the video stream of the respective clips for presentation. If more precision is needed in the alignment, a finer-grained binning could be applied, or a signal-level (or spectral-level) cross-correlation between the clips could be used.

4. NEW APPLICATIONS

While the synchronization data could directly be applied to the synchronized presentation of clips from a live show, we do not stop there. We apply the synchronization data in various ways that are useful for the application scenario of viewing and organizing the aggregate content from a live show. Next, we describe the various types of metadata, or knowledge, about the concert that we extract. First, though, we describe how we generate an implicit structure for the collection that underlies the metadata extraction.

4.1 Linking and Clustering

We use the audio fingerprinting matches to create a graph structure, linking between overlapping video clips. We use audio fingerprints to detect the fact that two video clips overlap temporally (i.e., both clips contain a capture of the same segment in the show). We treat each clip as a node in a graph, and create edges between each overlapping pair of videos. Thus, we generate a graph structure of video clips for the concert. Figure 4 shows the graph structure for one of the music concerts in our test data. The generated graphs tend to have a number of interesting aspects. First, the graph consists of several different connected components, which we sometimes refer to below as *clusters*: a collection of overlapping clips taken during the same portion of the show. Second, we observe that the connected components are typically not fully connected. In other words, not all clips in a cluster are detected as overlapping. This fact stems from two factors. First, overlap is not transitive. It is possible that a Clip A and Clip B overlap, Clip B and Clip C overlap, but yet Clip A and Clip C do not overlap (but are guaranteed to be taken in time proximity as clips are rarely more than few minutes long). More significantly, some connections between overlapping segments are not detected when matching fails, often due to a large amount of noise in the audio track. In Section 4.3 we show how these "missing links" can be helpful for improving audio quality.

It is possible that for a given set of clips, our system will generate a (undesirable) single connected component, but this scenario could be handled in various ways. For example, if one of the clips in our set is a continuous recording of the whole show or if there are many overlapping clips that essentially cover the whole show, the linking might result a single connected component. We did not encounter this in our experiments and our dataset. This problem might be easily dealt with by segmenting clips and only matching between smaller sub-clips, requiring stronger ties between sub-components, or using some graph partitioning approach, like normalized cuts [16].

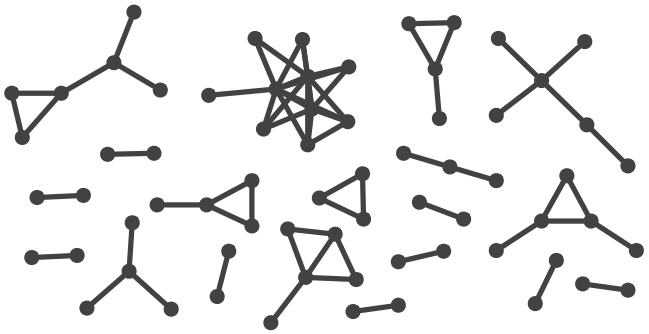

Figure 4: Example graph structure emerging from synchronizing and linking video clips from a concert based on overlapping audio content.

This emerging graph structure of clips can lead to some rich cues about the video clips and the event, and better representation and information for music event data in browsing community-contributed video collections on the Web. We discuss these effects and applications in the following sub-sections.

4.2 Extracting Level of Interest

The clip overlap structure, created by the community activity, can help identify moments in an event that are likely interesting to consumers of content [11]. In particular, we hypothesize 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 [1] or simple exploration of the event media. Videos of the most important segments or other aspects of the concert could be highlighted, while filtering lower-scoring clips that are either unrelated or, presumably, less interesting.

Our hypothesis is that larger clusters of matches between clips typically correspond to segments of the concert that are subjectively most "interesting." In the case of live music, these clusters could reflect significant moments in the show where a hit song is being played, or something particularly interesting is happening on stage. A very simple measure of ranking importance of clusters is simply counting the number of nodes (video clips) each overlap cluster contains. Alternatively, we can reason about the level of interest for any given clip by the clip's degree on connectivity. A highly connected clip is likely to depict a moment that many attendees of the show found interesting.

A segment of interest in a concert can range from a few seconds to a few minutes in length. The method that we propose does not have any length requirements for interesting segments, simply specifying the identified interest segment as the span of time covered by videos within the cluster. Alternatively, the "interestingness" analysis can be performed on sub-segments of the show and clips (e.g., 1 minute subclips) and interesting segments identified in that resolution.

4.3 Selection of Higher Quality Audio

We use the synchronization data to select the highest-quality audio for each overlapping segment. The synchronization between video clips can be used for playback, remixing or editing content. Such an environment could show all the available clips for a segment aligned with each other. Inevitably, given the nature of user-generated recordings, the video and audio quality and content can be highly variant between clips as well as from minute-to-minute within clips. For any remix or playback application, it may be desirable to select the highest-quality available audio, regardless of the audio source. The video playback, whatever it may be, can be overlaid on top of a the automatically-selected higher quality audio track.

Interestingly, low-quality audio tracks cause the audio fingerprinting method to fail in systematic ways that can be leveraged to point us towards higher-quality recordings. A weakness of the audio fingerprinting approach employed is that not all pairs of corresponding clips can be detected. The fingerprinting method works best when a database of completely clean sources is available and the objective is to match a very noisy copy to a source. As such, the system is highly robust against a remarkable amount of noise in one of the audio sources, as long as the other is relatively clean. However, if both sources are highly noisy, the probability of detection drops precipitously. In our data, we have many such cases of low quality capture. In these cases, our system will fail to detect matches between two low-quality recordings, but is more likely to detect matches between low-quality recordings and high-quality recordings. This situation is reflected in the resulting graph, and can lead us towards the highest-quality audio.

This observation yields a simple selection algorithm to identify higher-quality audio. Within a set of clips, which are all in the same cluster and are related to the same event segment, we choose the most-connected video clips as the probable highest-quality audio tracks.

4.4 Extracting Themes from Media Descriptions

We aggregate the textual information associated with the video clips based on the cluster structure to extract descriptive themes for each cluster. On many social media websites (including YouTube), users often provide lightweight annotations for the media in the form of titles, descriptions, or tags. Intuitively, if the overlapping videos within our discovered clusters are related, we expect the users to choose similar terms to annotate their 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. These terms are likely to be useful labels / descriptions for the cluster, and can also be used as suggested metadata for unannotated clips in that cluster.

We translate this intuition into a simple scoring method based on tf-idf (term frequency and inverse document frequency). We generate the set of words x that are associated with videos in a cluster C. We then score these terms using their cluster frequency, $tf_{C,x}$, or the number of videos for which the term x is used. We factor the score by general frequency of the term df_x, which is the count of video clips in the entire collection where the term x appears in the description. A score $s_{C,x}$ for each term is calculated as $s_{C,x} = \frac{tf_{C,x}}{df_x}$. The terms whose score exceeds an empirically-chosen threshold are selected to describe the content of the respective cluster.

4.5 Towards a Better Presentation of Concert Clips

A complete application for browsing event-based media is outside the scope of this paper, as we focus here on the algorithms and their direct evaluation. However, for completeness, we discuss below some possible features for such an application for viewing content from live concert videos. These features are based on the novel metadata we can generate, as described above.

The most straightforward application of our methods is the synchronized playback of clips. Figure 1 presents a possible viewing interface for synchronized playback. In the figure, we can see multiple concurrent videos being played in synchronized fashion. 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. Such automatic overlap detection could significantly reduce the required time to scan or watch all the content from a concert, while still providing a complete view. At any time, a single clip's audio track is played while the others are muted. Of course, the non-muted audio track can be selected based on its audio quality score, such that the best available can be playing with any user-selected video image.

The extracted terms representing each cluster of clip could be displayed during playback, but are more important for inter-cluster search and browse task. For example, a high-level view of all the content for a concert can show the major clusters, extracted keyframes from the clusters, and the key text terms that describe each as determined by our methods. The viewer could then quickly scan the available content and select to view the parts they are most interested in. Note that the extracted terms can also be used to suggest metadata to other clips in the same cluster that are not that well annotated.

The "interest" measure, which points out the most compelling clusters, can also be used to improve the browsing experience. Trivially, popular content (large clusters) could be featured prominently during browsing. Those clusters could also enjoy higher relevance score for a search scenario.

5. EVALUATION AND ANALYSIS

The assessment of different aspects of our application required diverse and often non-traditional evaluation techniques. We report below on our experience with audio fingerprinting in this application domain, including a discussion of the performance of the technique in matching clips with varied audio quality and without a 'master' audio source. We then describe our evaluation of the other aspects of our application, including detecting the event level of interest, selecting higher quality audio tracks, and extracting descriptive text for each set of matching clips.

First, however, we provide some descriptive data about our set of video clips. These clips were crawled and downloaded from YouTube and portray three live music events. The description of the dataset helps frame the subsequent evaluation, but is also a unique contribution as the first description (albeit limited) of the statistics of such a dataset.

5.1 Experimental Data

We have applied our system to a large set of real user-contributed videos from three concerts crawled from the popular video sharing website, YouTube. Each concert collection contains several hundred video clips, providing for a total of just over 600 clips and more than 23 hours of video footage. The three concerts that we have investigated are: Arcade Fire in Berkeley, CA; Daft Punk in Berkeley, CA; and Iron Maiden in Bangalore, India. All three concerts occurred during the spring or summer of 2007. As mentioned above, the details of the crawl exceed the scope of this paper. We do not have data regarding the coverage of these datasets. We do, however, believe that the set of clips for each concert is representative, and enjoys rather high recall.

	AFB	IMB	DPB
Total Clips	107	274	227
Precision	99%	97%	99%
Pre-/Post-Show	2%	10%	.5%
Opening Act	3%	11%	2%
Multiple Songs	1%	1%	6%
Synched	60%	25%	50%

Table 1: Summary of the characteristics of the online videos crawled for three concerts: Arcade Fire in Berkeley (AFB), Iron Maiden in Bangalore (IMB), and Daft Punk in Berkeley (DFB).

In Table 1 and Figure 5 we show a number of characteristics of the data sets that can yield some insights towards the way media is captured and shared, and suggest dependencies of the available data on the type of music being performed and the culture that the attendees share. Table 1 reports on a number of aspects of the recordings. First, the total clips line shows the total number of video clips found for each concert. The precision line reflects on the quality of the automatic video crawling system. Specifically, the precision is the percentage of the retrieved videos that were indeed captured at the concert. The false positives in the media crawl mostly consist of a clips of the same band performing in a different city, but sometimes include completely different of videos that have been mislabeled or miscrawled. These unrelated clips tend to make up a negligible portion of our data set, which suggests that the aggregation method is rather precise. The pre-/post- show line shows the proportion of clips that are not of the performance itself, but of social interactions between the concert-goers before and after the show, while the opening act line shows the proportion of clips that are of the opening act performing before the main show. In both of these cases, there are few such videos among the Daft Punk and Arcade Fire attendees, and a more significant portion from the Iron Maiden attendees. This fact may be a reflection on the cultural differences between these sets of attendees, or perhaps the relative infrequency of large rock concerts in Bangalore. Next, the table shows the percentage of multiple songs clips in which users edited together clips from different portions of the show into a single clip. This behavior is significantly more prevalent in the Daft Punk data, which may be influenced by the nature of the performance (discussed below). These multiple-song clips present a significant challenge for our system, since each of these clips may be linked to many different segments of the concert. Here, we manually remove these clips, though we believe that a mixture of shot-boundary detection and deeper constraints on the matching algorithm could result in a fully-automatic method for detecting such "re-mixed" clips. Finally, Table 1 shows the

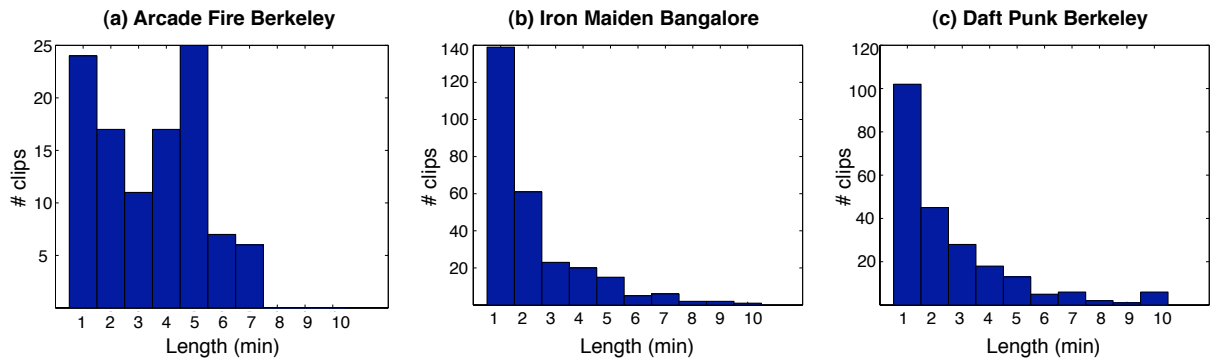

Figure 5: Distribution of clip lengths

number of synched clips, the proportion of clips for which the audio processing framework has found a matching clip (calculated using the optimal threshold as discussed next). This number can be negatively influenced by the presence of clips that are too short to synchronize (less than 10 seconds) or of extraordinarily low quality. It can also be influenced by the number of non-concert clips, where it is less probable that two users will capture the same content. Finally, the number of synchronized clips is of course a function of the density of available content: the more content is captured and shared, the more matches we shall find between clips.

The lengths of captured clips may also suggest, or be influenced, by characteristics of the show or of the attendees. Figure 5 shows the distribution of the lengths of video clips for each of the concerts. A number of interesting characteristics can be observed. In all cases, there is a heavy representation for clips shorter than a minute, with an immediate drop-off in the number of two or three minute-long clips. This drop is likely due to a number of factors, including the limitations of capture devices used (such as cellphones or digital still cameras) as well as the short attention spans typical in online video communities. There is a particularly strong bias towards the one-minute range in the Iron Maiden clips, which may be due to the larger prevalence of lower-end capture devices in India, particularly video-enabled cellphones. No clips are longer than 10 minutes, due to the limitations of the YouTube site. The type of music might also influence clip length. Two of the bands (Arcade Fire and Iron Maiden) are rock bands, and indeed, we observe a bump or leveling-out of the distribution in the range of four or five minutes – the length of a typical rock song – indicative of a tendency to capture and share entire songs. The Daft Punk performance does not consist of "songs" so much as a continuous mix of various elements of songs, so there is less of a tendency to capture entire songs. Conversely, there is a bump in Daft Punk clips at 10 minutes in length, the maximum YouTube clip length. This might be a reflection of the tendency observed above of attendees of this particular concert to construct summaries by using various clips, making them as long as the medium permits.

We executed the various algorithms on this dataset of videos from the three concerts. Next, we report the results.

5.2 Evaluation: Matching Clips

We evaluate the performance of the audio fingerprint-based method in finding matches between video clips. In particular, we wish to identify the correct threshold to be used for the audio fingerprinting match (Section 3.2), and to understand the trade-offs for different threshold values. We conduct this investigation by exploring the clusters resulting from various thresholds on the match detection score, as described in Section 4.1. Evaluating clustering results is inherently difficult. We take a pairwise approach: we look at each pair of video clips, and whether or not the ground truth indicates that they should end up in the same cluster. We evaluate whether or not the clustering results agree.

As discussed earlier, the clustering process is essentially conducted by forming disjoint sets through the discovered links between clips. At one extreme, with a very high threshold on the match score, the clips would all be completely unconnected as singleton clusters. At the other extreme, with a very low threshold, all of the clips would end up in the same cluster, which would lead to high pairwise recall, but of course, very low precision.

Our results suggest that the fingerprinting approach to linking can be highly precise, yielding very few errors in the final output. Figure 6 shows the precision-recall curves for each of the concerts in our data set, produced by varying the threshold and measuring, for each value, the recall and precision of matching at that point. The three curves all have virtually the same characteristics: near-perfect precision is maintained up to recall of 20% or 30%, after which point the precision drops off dramatically. It is important to note that the pairwise precision-recall evaluation that we are using is very aggressive. For example, a ground-truth cluster that is segmented into two clusters by the system results in a significant drop in recall. Similarly, two clusters that are wrongfully joined by the system can result in a large reduction in precision.

The price of precision is reduction in recall, or clips that will not be linked to others despite overlap – but this reduction may not be critical for our application. Inspection of the types of clips missed by the system indicates that many of these clips are frequently too short to synchronize (on the order of less than 10 seconds) or of abnormally low quality. Since our application is driven towards generating summaries and a better representation of an event, such clips would ultimately not be required. In fact, it may be better to be able to leave these clips out of any presentation. Furthermore, given the large number of clips, for each concert, it is typical to find about a dozen unique clusters, each with between 3 and 14 clips. Given the large number of available

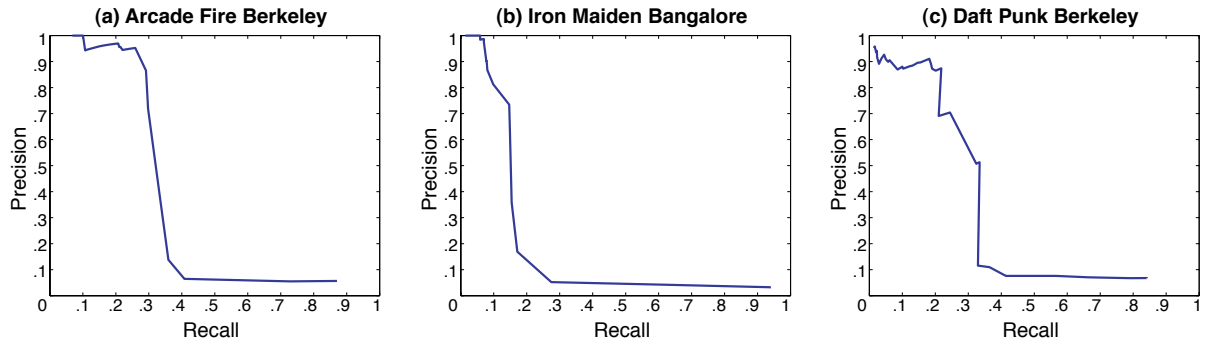

Figure 6: Precision (vertical axes) vs. Recall (horizontal axes) curves for pairwise detections of same-segment video clips.

relevant clips, we presume that the users are likely to prefer precision over recall. The remainder of the evaluations are conducted using a constant threshold: clips are considered to be in the same cluster if the largest bin of fingerprint offsets is seven standard deviations above the mean bin size.

5.3 Evaluation: Finding Important Segments

The goal of this part of the evaluation is to verify that our hypothesis and heuristics for finding subjectively important content are correct. In other words, we examine the resulting cluster structure for the three concerts, and determine whether the large clusters correspond to more significant occurrences, as described in Section 4.2. We first report qualitatively on the results of the importance measure. We then quantitatively compare our results to a large repository of user-contributed data of music interest.

Indeed, by inspecting the results of our clustering algorithm, we observe that the larger clusters typically correspond to segments of the concert that are subjectively most "interesting." In particular, we see that many of the large clusters resulting from the framework correspond to the most popular hit songs by that band. For example, the largest Daft Punk clusters contain videos of "One More Time," and "Around the World," which are two of the group's most popular singles. Similarly, the largest Iron Maiden clusters correspond to the songs "The Number of the Beast" and "The Trooper," which are (arguably) their most popular hits. In the Arcade Fire and Daft Punk concerts, nearly all of the clusters represent individual songs. In the Iron Maiden concert, nearly 20% of the segments correspond to non-song segments. Iron Maiden is well known for their stage theatrics, so examples of these non-song segments include parts where characters come out on stage or large props are introduced. These clips may also reflect segments where the singer is talking to the audience.

Our importance ranking proved to significantly match rankings found on the music sharing website Last.fm[1]. Of course, the determination of "interesting" or "important" segments of a concert is highly subjective and difficult to evaluate rigorously. To make another attempt at examining the validity of these results, we conduct a quantitative evaluation, comparing our ranking approach against some other metric of song popularity. For the purpose of this evaluation, we focus only on the segments where the target is a clearly identifiable song, and drop the non-song segments (like stage theatrics) from the evaluation. The external measure of song popularity is the data gathered from Last.fm, a popular social music website. In Last.fm, users can track which songs they are listening to and share this information with the community. One interesting aspect of Last.fm is its ability to trace the popularity of specific songs across the entire membership of the site. For each of the song-related clusters discovered by our algorithm, we identify the song manually, and compare the size of the cluster (the number of videos of the segment) against the number of times that the song has been played by Last.fm users. We find a statistically significant correlation between these two values ($r^2 \sim .44$, $p < .001$, $N = 41$). These results support the idea that the largest clusters indeed correspond to the most popular aspects of the concert.

It should be noted that despite the correlation, Last.fm cannot be used as a source of importance that replaces our method: first, the comparison was done under the assumption that the clusters are identified as songs, which is possible (see Section 5.5) but not guaranteed; second, preferred moments in live concerts might not map exactly to popular tracks. Finally, we should note that the raw popularity of recording a moment may not be the best indicator of interest or value in that moment. We might further augment this with how frequently users view the clips of a moment or how highly they rate those clips.

5.4 Evaluation: Higher Quality Audio

Can the cluster and link structure indeed suggest, as described in Section 4.3, which are the clips with better audio quality? To test that hypothesis, we conducted a study with human subjects. For the study, we took 15-second subsegments from each of the clips in the top two clusters for each concert, for a total of 6 clusters and 50 segments. The 15-second sub-segments for each cluster all correspond to the same segment of real-world audio such that their audio quality can be directly compared. For each sub-segment we had the quality score as generated by the system, which is the number of detected links to the respective clip. Is there a correlation between this score and a human rating?

We asked two independent human subjects to listen to the recordings in a random order and to score the clips' quality on a scale of 1 (unintelligible) to 5 (professional quality). As a sanity check, we first compared the scores provided by the two subjects against each other and found a significant correlation ($r^2 \sim .58$, $p < .001$, $N = 50$), which suggests

[1]http://www.last.fm

Show	Automatic Text	Content Description
AFB	go cars no	song: "No Cars Go"
	headlights	song: "Headlights Look Like Diamonds"
	intervention	song: "Intervention"
	wake	song: "Wake Up"
	keep car running	song: "Keep the Car Running"
	tunnels	song: "Neighborhood #1 (Tunnels)"
	lies rebellion	song: "Rebellion (Lies)"
	back seat	song: "In the Back Seat"
	laika	song: "Laika"
	cold fire	song: "Cold Wind"
IMB	trooper	song: "The Trooper"
	greater god	song: "For the Greater Good of God"
	hallowed be thy name	song: "Hallowed Be Thy Name"
	number of the beast	song: "The Number of the Beast"
	bruce dickinson	[singer, Bruce Dickinson, introduces band]
	frontman	song: "Wrathchild"
	evil eddie	song: "The Evil that Men Do," [Eddie, band mascot, enters]
	fear of the dark	song: "Fear of the Dark"
	tank me	[Eddie, band mascot, appears in a tank]
DPB	stronger faster around	song: "Harder, Better, Faster, Stronger" mixed with song: "Around the World"
	encore	[show encore]
	da funk	song: "Da Funk"
	live	song: "Steam Machine"
	face	song: "Face to Face"
	intro	[introduction to concert]
	ca	song: "Technologic"

Table 2: Examples of automatically-generated text labels (generated by proposed algorithm) for clusters found for Arcade Fire in Berkeley (AFB), Iron Maiden in Bangalore (IMB) concerts and Daft Punk in Berkeley (DFB), along with the corresponding songs and, when relevant, brief descriptions of the actual occurrences (both provided manually by the authors).

that this audio-quality identification task can be reliably done by humans. We then averaged the scores from the two subjects and compared that value against our audio-quality scores (i.e., the number of links, or detected overlapping clips, for each given clip). We found a significant correlation ($r^2 \sim .26$, $p < .001$, $N = 50$), which suggests that our proposed approach is indeed successfully detecting higher-quality audio segments.

5.5 Evaluation: Extracting Textual Themes

Perhaps the most difficult aspect of the system to evaluate is the quality of the text that is automatically associated with each of the discovered clusters. The "correct" result is subjective and a human-centered evaluation would require subjects that are intimately familiar with the artist's songs and might even require subjects who physically attended the concert being explored. Instead of conducting a numerical evaluation of these results, we simply list most of them in Table 2 along with some intuitive descriptions of the actual content of the videos to provide the reader with a qualitative notion of the type of output provided by this system.

By scanning through the table we can see that, in many cases, the text keywords extracted by the system for each cluster indeed mirror some important keywords from the title of a song or describe some aspect of the stage actions. In particular, a few song titles, such as "Intervention," "Laika," "Hallowed Be Thy Name," "Fear of the Dark," and "Da Funk" are recovered perfectly. Other titles, like "Rebellion (Lies)" or "Harder, Better, Faster, Stronger / Around the World," for which the system returns "lies rebellion" and "harder stronger around," respectively, the ordering of terms is mixed up and some terms may even be missing. Nonetheless, these annotations may still provide helpful clues about the video cluster to users who are knowledgeable about the artist. Still other categories are more reflective of the moments happening on stage, rather than the titles of songs. For example, the Daft Punk clusters corresponding to the introduction and encore portions of the concert were automatically labeled "intro" and "encore," respectively. Similarly, the Iron Maiden clusters corresponding to stage theatrics portions of the concert where the band mascot, Eddie, comes on stage (first on foot, and then in a tank) are labeled "evil eddie" and "tank me," respectively, both solid clues regarding the content of each cluster. The term "frontman" in Table 2 describes the cluster of videos of Iron Maiden's performance of "Wrathchild." Manual examination of the videos associated with this generated description reveals that the lead singer chose to dedicate the performance to a former (deceased) frontman for the group, a fact that most users chose to denote in their textual description, and therefore was reflected correctly in our results.

Finally, there are of course some failure cases. In particular, the performances by Daft Punk of "Steam Machine" and "Technologic" are inappropriately labeled "live" and "ca." These tags are particularly unhelpful, since all the clips are of a live show in California. In these cases, manually inspecting the clips in the cluster, we found few useful user-provided text associated with the original context. The failure cases that we observe tend to correspond with smaller clusters (of only 3 or 4 videos), where the volume of videos does not outweigh the general aversion that users show towards providing detailed annotations. Such low availability of textual content might be relieved by including text from websites that embed the video clips, beyond the YouTube site.

By and large, we see that a great deal of information about the content of video clusters can be revealed through mining the textual notations attached to the videos within the clusters. In particular, we find a great deal of keywords sampled from the song titles, and sometimes even fully-formed titles. This type of information can provide much-improved browsing mechanisms for the content of a concert, with no explicit editorial or curation process.

6. CONCLUSIONS AND FUTURE WORK

We have set the grounds for an application that enhances the presentation and findability of content in live music concerts, a popular multimedia application scenario on the Web. The ideas above could be used in the design and implementation of a system for sharing live concert videos and content. We would also imagine such an application to elicit more accurate or structured metadata and contributions from users,

contributions that might exceed and extend the social media tools available on YouTube. Such an environment could even further enhance and improve the consumption of Web-based multimedia associated with the event experience.

Some key problems remain in working towards that system. We did not provide the details of the crawling algorithm that aggregates video clips from one show. While we do have an initial algorithmic solution to this problem, a refinement of that work and a human-centered approach could help the accuracy and recall of a crawler. Of course, the design of such a system as an interactive, lively environment is a challenge, and so is incorporating different content analysis tools and results.

Different application scenarios could impact the design of the system and the available features. In particular, what can we accomplish if there is an authoritative source of audio (or audio and video) from the event? Such professionally-produced content mixed with social media contributions can significantly change the viewing experience [10] as well as the requirements from our system in terms of algorithms and processing.

Generally, our ideas in this work serve as an example for fusion of context provided by social media sites, and content analysis (our work in [8] provides another example for this approach in a different domain): we use the new context available from social media sites like Flickr and YouTube to reduce the complexity and the required scale of the content analysis tasks. Furthermore, we use the aggregate results of the content-based match in conjunction with metadata from social media with the to create a better representation of the content. In the case of this work, audio fingerprinting features can be extracted only for clips identified as events using the social media context, and clips are pairwise compared only within a single show, significantly reducing the required scale as well as potentially improving precision. This approach for multimedia analysis leveraging social media contribution promises to change the way we consume and share media online.

7. REFERENCES

[1] M. G. Christel, A. G. Hauptmann, and H. D. Wactlar. Collages as dynamic summaries for news video. In *MULTIMEDIA '02: Proceedings of the 10th international conference on Multimedia*, pages 561–569. ACM Press, 2002.

[2] S. J. Cunningham and D. M. Nichols. How people find videos. In *JCDL '08: Proceedings of the Eigth ACM/IEEE joint conference on Digital libraries*, New York, NY, USA, 2008. ACM.

[3] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd. Time as essence for photo browsing through personal digital libraries. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, 2002.

[4] J. Haitsma and T. Kalker. A Highly Robust Audio Fingerprinting System With an Efficient Search Strategy. *Journal of New Music Research*, 32(2):211–221, 2003.

[5] C. Herley. Accurate repeat finding and object skipping using fingerprints. In *MULTIMEDIA '05: Proceedings of the 13th international conference on Multimedia*, pages 656–665. ACM Press, 2005.

[6] C. Herley. ARGOS: automatically extracting repeating objects from multimedia streams. *IEEE Transactions on Multimedia*, 8(1):115–129, 2006.

[7] V. Kaplun, P. Vora, M. Naaman, P. Mead, and A. Moed. Understanding media capture and consumption for live music events. Technical report, Yahoo! Inc., 2008. In Submission.

[8] L. Kennedy and M. Naaman. Generating diverse and representative image search results for landmark queries. In *Proceedings of the Seventeenth International World Wide Web Conference*, New York, NY, USA, 2008. ACM.

[9] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries*, 2004.

[10] S. U. Naci and A. Hanjalic. Intelligent browsing of concert videos. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 150–151, New York, NY, USA, 2007. ACM.

[11] R. Nair, N. Reid, and M. Davis. Photo LOI: Browsing multi-user photo collections. In *Proceedings of the 13th International Conference on Multimedia (MM2005)*. ACM Press, 2005.

[12] J. Ogle and D. Ellis. Fingerprinting to Identify Repeated Sound Events in Long-Duration Personal Audio Recordings. *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 1, 2007.

[13] H. W. Prarthana Shrestha, Mauro Barbieri. Synchronization of multi-camera video recordings based on audio. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 545–548. ACM Press, 2007.

[14] D. Shamma, R. Shaw, P. Shafton, and Y. Liu. Watch what I watch: using community activity to understand content. *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 275–284, 2007.

[15] R. Shaw and P. Schmitz. Community annotation and remix: a research platform and pilot deployment. In *HCM '06: Proceedings of the 1st ACM international workshop on Human-centered multimedia*, pages 89–98, New York, NY, USA, 2006. ACM.

[16] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 888–905, 2000.

[17] P. Shrestha, H. Weda, M. Barbieri, and D. Sekulovski. Synchronization of multiple video recordings based on still camera flashes. In *MULTIMEDIA '06: Proceedings of the 14th international conference on Multimedia*, pages 137–140. ACM Press, 2006.

[18] C. Snoek, M. Worring, A. Smeulders, and B. Freiburg. The role of visual content and style for concert video indexing. *Multimedia and Expo, 2007 IEEE International Conference on*, pages 252–255, 2-5 July 2007.

[19] S. Uchihashi, J. Foote, and A. Girgensohn. Video manga: generating semantically meaningful video summaries. In *MULTIMEDIA '99: Proceedings of the 7th international conference on Multimedia*, pages 383–392. ACM Press, 1999.

[20] Y. van Houten, U. Naci, B. Freiburg, R. Eggermont, S. Schuurman, D. Hollander, J. Reitsma, M. Markslag, J. Kniest, M. Veenstra, and A. Hanjalic. The multimedian concert-video browser. *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1561–1564, 6-6 July 2005.

[21] A. Wang. An Industrial Strength Audio Search Algorithm. In *Proceedings of the International Conference on Music Information Retrieval*, 2003.

[22] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE Multimedia*, 14(1):19–29, 2007.

[23] Youtube.com, google inc. http://www.youtube.com.

[24] A. Zunjarwad, H. Sundaram, and L. Xie. Contextual wisdom: social relations and correlations for multimedia event annotation. *Proceedings of the 15th international conference on Multimedia*, pages 615–624, 2007.