

# If It Sounds As Good As It Looks: Lessons Learned From Video Retrieval Evaluation

**Abby A. Goodrum**

School of Information Studies

Syracuse University

Syracuse, NY 13244-4100

+1 (315) 443-5602

aagoodru@syr.edu

## ABSTRACT

In many ways, music information retrieval (MIR) bears a closer resemblance to video information retrieval (VIR) than it does to text retrieval. Both music and video provide rich, complex sources of information having their own semantics. Both share the challenges of digitizing, segmenting and streaming, joined by problems relating to the representation of non-textual, non-verbal information. Because of this complexity, systems for the retrieval of these media pose unique challenges to evaluation including the construction of large testbeds, the crafting of representative topics for searching, and identification of appropriate metrics for evaluation. This paper will discuss recent efforts in video retrieval evaluation and how these efforts might inform the creation of an experimental evaluation environment for Music Information Retrieval.

## 1. INTRODUCTION

Evaluation is the process of examining an entity (subject, process, system, technique, etc.) and assessing or appraising it based on its important features. We determine how much or how little we value an entity, arriving at our judgment on the basis of criteria that we can define and measure. As such, evaluation is a critical component in the progress of science and technology.

The field of IR has a long tradition of evaluation used to compare the relative performance of different system designs. For the past 12 years, the National Institute of Standards and Technology (NIST), part of the US Department of Commerce, has been running a series of conferences on text retrieval (TREC). TREC has gathered together large collections of text, spoken audio, web and video information with a view to supporting research into information retrieval (IR). TREC provides an infrastructure and mechanisms for the comparative evaluation of IR systems that have greatly increased our understanding of IR over the last decade. This is no ‘vulgar pick the winners approach,’ [12] but is instead an attempt to uncover the best solutions to support the information needs of users.

All evaluative endeavors require the identification of suitable criteria for evaluation, measures and instruments for measuring the criteria, and methodologies for conducting evaluation experiments. Several scholars have pointed to problematic issues within IR evaluation [4,7,8,13,14], and I will not attempt a complete recitation of those issues here. I wish to focus, instead, on those evaluation issues which are shared by video information retrieval and music information retrieval. I do not claim that the issues discussed here are not pertinent for text information retrieval researchers, but the focus will be on how these challenges relate specifically to video and music IR. Specifically, I will discuss the unique methodological problems of constructing large media test collections, crafting statements of information need, and selecting appropriate mechanisms for analysis of results.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

As a first step, it is important to examine the unique characteristics of music and video information sources that distinguish them from textual information sources.

## 2. SHARED CHARACTERISTICS

Digitized music and video are rich and complex sources of information. While they may convey information similar to that contained in texts however, they do not communicate in the same fashion as texts. Moreover, both music and video impart information sequentially over time at varying levels of complexity.

For example, video is composed of approximately 25-30 individual images or frames per second. Each frame contains information relating to both low-level and high level features including shapes, textures, colors, brightness, and the position or location of complex objects such as people, places, and things. An unbroken sequential string of frames taken from the same camera defines a shot and includes transitions such as fades, hard cuts, dissolves, and wipes. Successful shot boundary detection paired with feature extraction is a prelude to scene detection, wherein an object of interest is visible across multiple shots that do not necessarily occur contiguously. Correctly identifying high-level video structures is a difficult task hence automatic detection of fundamental units, such as shots, from the stream is vitally important.

Similarly, music information contains both high and low-level features such as notes, dynamics, intervals, key, loudness, timbre, melody, rhythm, pitch, voice, instruments, timing, noise, tonality, themes, etc.. This is complex enough when discussing monophonic music (one note occurring at a time), but the level of complexity increases substantially when considering polyphonic music that may be comprised of multiple voices and instruments simultaneously. Identifying appropriate mechanisms for feature extraction and segmentation are crucial here as in video.

In addition to having complex data structures, digitized music and video have file structures and file sizes that impact compression, storage and retrieval. Uncompressed music files are large, take up memory, and are slow to search and download. Compression makes them easier to transmit and store, but results in some loss; typically inaudible of redundant data. For example, one minute of CD quality music in mp3 format is roughly equivalent to 1 MB. To determine the file size of one second of uncompressed video, multiply the image size by the number of frames per second (fps). For example, one second of uncompressed, full-size, full-speed (30 fps), 24-bit video is:  $900K \times 30 = 27 \text{ MB}$  or 2.7 MB using a compression ratio of 10:1.

The challenges of digitizing and segmenting are joined by problems relating to the representation of non-textual, non-verbal information. Finding, for example, all instances of a certain pitch, harmony, rhythm, shape, texture, or visual object challenges IR systems built to essentially match words in a query to words in a collection. The problem is essentially one of representational congruity.

Representation is a central concept in information retrieval, and occurs at several stages in the process. On one level, authors/composers/filmmakers represent their ideas and knowledge as documents such as articles, books, scores, and films. Similarly, users' represent complex goals, problems and knowledge gaps as information needs. At another level, representations of documents are matched against representations of the users' information need as expressed in queries. Successful retrieval is predicated on the extent to which representations actually share in the nature of the thing being represented. There are three areas that affect the outcome of the IR system with respect to representation:

- ✍ The extent to which document representations share congruence with the documents for which they stand.
- ✍ The extent to which queries share congruence with the information needs for which they stand.
- ✍ The extent to which queries and document representations share congruence with each other.

Representations for documents function not only as attributes against which a query may be matched, but also provide support for browsing, navigation, relevance judgments, and query reformulation. It is important to note, however, that the representation of a users' information need as expressed in a query is largely driven by system parameters for query construction and for representation of retrieved documents. Only in the last decade have users been able to query by humming or query by submitting an image exemplar. Thus, recent advances in the technology for signal processing and pattern matching have been driving decisions about MIR and VIR system design rather than an understanding of user needs and user behaviors in these retrieval environments. This is not to suggest that we embrace an either-or scenario for video and music IR research. Both system and user-centered approaches are needed at this early stage.

Both video and music are multidimensional and require multiple features in order to represent a 'document,' but there is sparse research to drive our understanding of which features are most useful for searching, sorting, ranking, navigating and relevance judgments. There has been a great deal of research demonstrating that relevance is a complex, multi-faceted relationship among information needs, information tasks, information environments, documents, document attributes, search processes and search interfaces [9]. Increasingly we are coming to understand that the criteria for relevance must be defined within a specific task context

before evaluation can occur. For video and music IR we lack a taxonomy of tasks and a taxonomy of queries related to these tasks. For appropriate evaluation design, we must also identify the various features that support relevance judgments for those tasks.

Relevance is also central to the two measures used in most IR evaluation studies: recall and precision. Recall is the proportion of relevant document that are retrieved, and precision is the proportion of retrieved documents that are relevant. In VIR and MIR these measures raise a number of questions including, (but not limited to):

- ✍ What constitutes a document for the purposes of computing precision and recall?
- ✍ What portion of the document constitutes an 'answer'?
- ✍ Is the entire video or score relevant to an information need, or only that segment containing the feature of interest?
- ✍ How do we assess the performance and contribution of different types of features?

Given the sparsity of empirical understanding in these areas, it would seem daunting to attempt any large scale comparative evaluation across systems. Nevertheless, in an effort to move VIR research forward, the VIR community began planning for and developing what later became the video track at TREC. A brief overview of the evolution of this endeavor may provide a framework for discussion of ways that the MIR community can prepare to move in a similar direction.

### 3. TREC VIDEO TRACK

For many years, VIR researchers constructed their own testbeds of digitized video to support their projects. These proprietary collections in many cases contained copyrighted material, or material lacking documented intellectual property rights. For example, the Informedia Project at Carnegie Mellon built over a terabyte digital video library of data from CNN, The Discovery Channel, and public television [16]. Each VIR project also created a variety of surrogates and indexing schema to support the focus of their particular research goals.

By the late 1990's the video information retrieval community was beginning to call for a large-scale open test environment for evaluating the various content-based retrieval systems that had emerged from research labs around the world. One of the central challenges in this was the creation of a large collection of freely available, public domain digitized video covering a diverse range of topics and representative of real retrieval environments.

At the Twenty-Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, in Berkeley, California, the National Institute of Standards and Technology (NIST) publicized the release of the first installment of a public domain digital video test collection on DVD [10]. At about the same time, the Open Video Project at the University of North Carolina also made available a large collection of digitized video for use by the VIR community [5,11].

Starting in 2001, TREC sponsored the first research track devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. Currently in its third year, the goal of the TREC video track has been to promote progress in content based retrieval from digital video via open, metrics based evaluation [15]. Beginning in 2003, this track will become an independent 2day evaluation workshop taking place each year just before TREC.

From the start, the intent of the TREC video track was to allow the methods and procedures for evaluation to evolve from year to year, based on participant feedback. In two short years, the video track participants have made substantial progress in creating mechanisms for the evaluation of video retrieval systems. The amount of digitized video used for the test collection has grown from 11 hours of video data in 2001 to approximately 73 hours of publicly available VCD/MPEG-1 from NIST, the Internet Archive and the Open Video Project. The subject material in the test collection has also been expanded and diversified and includes both black & white and color films produced between the 1930s and 1970's. The subject matter now ranges across topics taken from educational, industrial, amateur, public service, and marketing films. The content of the topic collection and the process for determining relevance have also evolved substantially thanks to the combined efforts of the video track participants and other members of the video retrieval community. What follows is a brief overview of the evolution of the TREC video track over the past two years.

#### *Tasks*

The tasks have changed from the first year. In 2001 the tasks were shot boundary detection, known item search, and general search. Both search tasks could be conducted in either automatic or interactive mode, and systems were allowed to use transcripts created from automatic speech recognition. Participants in the 2002 video track took part in three tasks: shot boundary detection, feature extraction (new in 2002), and searching. The searching task was modified after 2001 to exclude fully automatic topic-to-query translation, and the known item and topic searching has been discontinued.

#### *Shot Boundary Detection*

Although single keyframes have been used successfully as surrogates for video retrieval and relevance judgments, an

important challenge in video processing for many years has been the ability to automatically discern different types of shots. A shot is defined as an unbroken sequence of frames taken from one camera and includes shot transitions such as fades, hard cuts, dissolves, and wipes. In the camera, a shot is what is recorded between the time the camera starts rolling and the time it stops rolling. In collections that use shot-level description, such as stockshot libraries or television archives, there may be an initial edit of the raw material, for example to discard technically imperfect or unusable parts of a shot. As a result, the shot found in the information system may be different from the one that originally came from the camera.

Data for the shot boundary task in 2001 was comprised of 5.8 hours (3.34 gigabytes) of MPEG-1 encoded video. In 2002 this was cut to 4 hours and 51 minutes of MPEG-1 encoded video. All transitions were identified and classified beforehand by NIST. In the course of running the shot boundary task in 2001, participants discovered that different MPEG-1 decoders were producing varying frame numbering from the same source video files. Work around solutions were proposed by participants and modifications were made to the protocols used for comparing submissions against reference data.

Shot detection is also an important prelude to the evaluation of search performance as well. Search runs typically point to 'answers' found within multiple shots that must be well-defined and robust in order to support comparison against the search runs of competing systems. Problems arose in 2001 wherein the shot boundaries for search tasks were defined by participants making comparison across systems difficult. For this reason, predetermined shot definitions were used by all groups in 2002 for the evaluation of feature detection and search tasks.

#### *Feature Extraction*

The ability to automatically identify the presence of high-level features such as "People", "Indoor/Outdoor", "Text Overlay", and "Instrumental Music" etc. represents a significant achievement in video retrieval. Additionally, these features may serve as a basis for enhanced search capabilities

The semantic feature extraction task was introduced in 2002 and will be continued in 2003. The objectives of this task were to begin the benchmarking process for evaluating effectiveness of detection methods on different features, and to allow for features extracted from this task to be available to participants in the search task as part of queries. During on-line discussions by track participants, a simple set of features were chosen that had the greatest potential given current system abilities. These features were suggested by participants before the topics for search were known.

#### *Topics*

Ideally, topics should be taken from real users interacting with the same collection used in the TREC video track. This was not an option however, so statements of information need (topics) in the first year of the TREC video track were created by participants and by NIST. NIST was also responsible for making some revisions and eliminating duplicate topics. Topics were subdivided into known-item and general searches. All topics were pooled and all systems were expected to run on all topics. Each topic statement included a textual description of the information needed and one or more media exemplars (audio, video, still image). Relevance of both known item search topics and general search topics was assessed by NIST.

For the TREC video track in 2002, 25 topics were created by NIST to represent the needs of a trained user seeking material for reuse in a large video archive. Known item topics were discontinued. As before, the statements included multimedia exemplars. An as yet unsolved problem remains: the topic statements are created from observation of the collection and may be biased toward text or audio accompanying some of the videos.

#### *Search*

Research examining human video retrieval interaction is still in its infancy, and documentation of how users formulate and modify queries using multiple media is scant. Moreover, a taxonomy of actual query types across diverse disciplines, collections, user characteristics, etc. has yet to be developed.

From the start, the TREC video track community recognized the need to incorporate human cognition into the search task. Searching is subdivided into two approaches: a 'manual' approach whereby a human searcher formulates a single query optimized for a specific search system based on the topic description. And an 'interactive' run in which a human searcher generates an initial query and then refines that query based on initial search output. In addition to calculating mean average precision, the interactive runs also measure total elapsed time for each search.

Although the first TREC video track in 2001 provided for the evaluation of fully automated search runs, this has been set aside for the time being and will be revisited in a future workshop.

Systems may be developed with knowledge of the test collection beforehand groups may also take advantage of the features donated by groups participating in the feature extraction task. The diversity of searcher and search interface interactions as well as the variability across topics makes a comparison across systems somewhat difficult and a goal for 2003 is to explore methods to improve this state of affairs.

Given that the TREC video track is still evolving and defining the methods, data, parameters, criteria, measures and procedures for

conducting comparative evaluation across VIR systems, what can the MIR community learn from their experiences?

#### 4. TREC MUSIC TRACK?

A diverse range of systems have so far been developed for music information retrieval, including systems that retrieve from MIDI representations, monophonic transcriptions, scores, thematic catalogs, and raw audio files [2,6]. A preliminary question to ask is whether a sufficient number of systems having comparable approaches exist. Furthermore, would the creators of these systems be interested in participating in an evaluative study?

At the same time, research examining music information behaviors is sparse [1]. A second preliminary question to ask is whether enough is known about music information seekers/users to begin to create topic statements and relevance assessments. Task definition is vital: who are your users and what do they need to do? Knowing this makes a difference in the type of collection you create, the types of queries you support, and the nature of relevance judgments. Given affirmative answers to the above and drawing upon the experiences of the TREC video track participants, what should the MIR community do next?

The first effort should be the creation of a large, easily accessible collection of music. This testbed collection for MIR should reflect the diversity of real collections. It should contain recordings of various length, from a range of genres, artists, and instruments. There should be multiple recordings of the same composition as well as variations on themes. The recordings should be supported by additional resources such as commentary, critique, indexing, scores, annotations, and other metadata describing the files. The collection must be free of intellectual property usage restrictions.

Next, the MIR community must develop a test collection of complex and diverse queries – including humming by real people, real queries, known item queries, general pattern matching and genre queries, etc. If possible, these queries should be generated by real users interacting with real collections and their relevance judgments should be captured along with their search processes.

Finally, don't try to do it all in the first year! The video track in TREC is still evolving and will continue to do so for many years to come. Similarly, the complex issues surrounding music information retrieval evaluation will not be solved or settled in a single workshop. That should not stop the MIR community from moving forward – with caution and an eye toward iterative improvement and gradual understanding of the complexities of evaluation in multimedia environments.

#### 5. CONCLUSIONS

Content based video IR and music IR are in their infancy. We have only recently moved from a bibliographic paradigm rooted in text retrieval to the development of retrieval systems based on visual and audio feature matching. Furthermore, we are still a long way from developing systems that are capable of human-like understanding of audio and video. An important step towards this understanding will be to expand our knowledge of user interactions with these media and to incorporate this knowledge into the design and evaluation of systems for retrieval.

While a TREC music information retrieval track is a good idea, we should remember that this is not the only way (and possibly not the best way) to evaluate MIR systems. Other approaches including case studies of MIR systems use in multiple organizational settings, qualitative studies of MIR system users and of music information seeking behavior outside of specific systems will also yield useful insights. In addition to precision and recall, other metrics should be explored such as those relating to browsing, navigation, and music understanding. Similarly, evaluation should be focused at different levels of feature complexity.

Finally, although we certainly wish to conduct comparative evaluation on the newer content-based technologies for music information retrieval, we should not abandon older text-based approaches entirely. A great deal of music is still cataloged and indexed in this fashion and will be for many years to come. Moreover, inclusion of bibliographic approaches allows for wider participation in MIR system research and development by a more interdisciplinary MIR community [3].

#### 6. REFERENCES

- [1] Cunningham, S.J. (2002). User studies: A first step in designing an MIR testbed. Papers Presented at the Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation, 18 July, 2002. pp. 19-21
- [2] Downie, J.S. (2000) Access to Music Information: The State of the Art. Bulletin of the American Association of Information Science & Technology. Vol 26(5), June/July 2000.
- [3] Futrelle, J. & Downie, J.S. (2002). Interdisciplinary communities and research issues in music information retrieval. Proceedings of the Third International Conference on Music Information Retrieval: ISMIR 2002 Paris, France, October 13-17, 2002, pp. 215-221
- [4] Ingwerson, P. (1992). Information retrieval interaction. London: Taylor Graham.
- [5] Open Video Project <<http://openvideo.dsi.internet2.edu/>>

- [6] Reiss, J.D. & Sandler, M.D. (2002). Benchmarking music information retrieval systems. JCDL Workshop on the Creation of Standardized Test Collections, Tasks and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation. Portland, Oregon, 2002.
- [7] Robertson, S.E. & Hancock-Beaulieu, M.M. (1992). On the evaluation of the IR systems, *IP&M* 28, (4), 457-466.
- [8] Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *Proceedings of ACM SIGIR '95*, Seattle, WA. Pp. 138-146.
- [9] Schamber, L. (1994). Relevance and information behavior. In Williams, M. (Ed.) *Annual Review of Information Science & Technology*, vol 29, (pp3-48). Medford, NJ: Learned Information.
- [10] Schmidt, C. & Over, P. (1999, August). Digital Video Test Collection. In proceedings of the Twenty-Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, California, USA.
- [11] Slaughter, L., Marchionini, G., & Geisler, G. (2000). Open video: A framework for a test collection. *Journal of Network and Computer Applications*, 23.
- [12] Spark Jones, K. (1995). Reflections on TREC. *Information Processing & Management*, 31 (3). 291-314.
- [13] Spark Jones, K. (Ed.) (1981). *Information Retrieval Experiment*. London: Butterworths.
- [14] Tague-Sutcliffe, JM.. (1992). The pragmatics of information retrieval experimentation, revisited. *IP&M* 28(4), 467-490.
- [15] TRECVID website:< <http://www-nlpir.nist.gov/projects/trecvid/>
- [16] Wactlar, H., Hauptmann, A., Gong, Y., Christel, M., (1999). Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library *IEEE Computer* 32(2): 66-73.