Mining Cinematic Knowledge: Work in Progress

[An Extended Abstract]

Duminda Wijesekera*
Department of Information and Software Engineering
George Mason University, MS 4A4,
Fairfax, VA 22101, U.S.A.
duminda@ise.gmu.edu

Daniel Barbará†
Department of Information and Software Engineering
George Mason University, MS 4A4,
Fairfax, VA 22101, U.S.A.
dbarbara@ise.gmu.edu

ABSTRACT

This paper presents the blueprint of an on going effort underway at George Mason University to create a movie mining system that uses already existing content detection technology. The emphasis of this project is to examine the suitability of existing concepts in data mining to multimedia, where the semantic content is time sensitive and constructed by fusing data obtained from component streams. Methods used for this purpose have a non zero probability of being incorrect. We discussed the issues involved in mining knowledge from or about movies and propose some solutions.

Categories and Subject Descriptors

data mining [**multimedia**]: cinematic knowledge

Keywords

data mining,multimedia,cinematics

1. INTRODUCTION

Multimedia and data mining are two young and flourishing fields, with their own application domains: multimedia concentrating on video conferencing, VoD services, databases, content based retrieval and data mining concentrating on detecting *interesting* patterns from basket data, event scripts, web-based information etc.

In detecting *interesting patterns* contained in multimedia data, one of the basic problems that needs to be addressed is the issue of extracting semantic information from audio, images, and video - and this has proven to be a challenging problem with limited success in specific applications. Notice that this problem does not arise in conventional data types. Secondarily, even when it is possible to extract and track basic features such as faces and gestures from surveillance video, it is still difficult to translate those gestures and facial movements to semantic information. There are many problems involved in doing this First, with the current state of the art in image and audio, it is not easy to identify *semantic information* easily. Secondly, higher *semantic* content may be contained in more than one stream of data, and consequently, some mechanisms to *fuse* these streams in order to extract the common semantics are needed. The third problem is the suitability of applying existing data mining concepts and algorithms (that have been successful with traditional textual data) to complex, fused multimedia data with different *quality of service* characteristics - which can be answered only when there is a measure of success in addressing the first two issues. Fortunately, as evidenced from work in image and audio analysis [12, 27, 31, 38, 11, 29, 32, 26, 13] [8, 39, 4, 16, 3, 7] there are tools to begin mining knowledge contained in multimedia data.

The objective of this extended abstract is to present a project that is underway at George Mason University where such audio and video analysis techniques are being used to mine *interesting knowledge* from multimedia data contained in movies. There are many reasons for selecting cinema. First, except for the early *silent movies*, cinema involves more than one media stream, and consequently is a good place to begin mining knowledge from data *fused* from multiple media streams. Secondly, movies on compact disks and tapes are relatively easy to access. In the third place, modulo some differences of opinion, movies have some *structure*, such as scenes and shots etc, and there are some standard cinematic styles. Further, at least the structural boundaries of the former can be detected with existing technology [39, 4, 31] (subject to a small amount of uncertainty.) Fourthly, mined knowledge can be verified against human audiences, thereby gaining a measure of validation of the correctness and appropriateness of mined knowledge. Considering all these facts, Cinema serves as a good test case for mining for knowledge contained in more than one media type, and consequently to look for appropriateness of existing mining algorithms and concepts for multimedia.

The rest of the paper is organized as follows. Section 2 de-

*Partly supported by an Internal Grant from George Mason University

†Partially supported by NSF grant IIS-9732113

© The copyright of this paper belongs to the paper's authors. Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage.
Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

scribes relevant background in the area of multimedia data mining. Section 3 describes our work in progress, including our objectives and plans to achieve them. Section 4 describes the prototype testbed that is being constructed. Section 5 contains concluding observations.

2. BACKGROUND

This section provides a brief summary of techniques and tools that are relevant to our work. A longer summary of work in multimedia data mining appears in [36].

The *Automatic Movie Content Analysis (MoCA)*[12] project at the University of Mannheim has developed or improved a lot of techniques to segment video and audio. They have also developed other algorithms such as for face detection, genre recognition of movies, recognizing text and music in TV commercials etc. Most of these techniques can be used to mine interesting patterns hidden in movies, or portions thereof.

One of the main issues that have to be recognized in mining for structure in movies is recognizing their boundaries. There are many methods to recognized scene cuts using the video [39, 4], audio [7, 38, 9] or both [31]. With the use of commercial tools such as the speech-to-text translator Dragon [11] and text analyzers such as WordNet [3], it is possible to enhance the detection of structure in movies.

The work at IBM Almaden Research Center covering *Query by Image Content (QBIC)*[13] and *QueVideo* [8] projects have made significant advances in recognition, indexing and content based retrieval of video [33], audio [34] and images. Similarly, the *Informedia* Digital Libraries project at Carnegie Mellon University [7] has made significant advances in indexing and analysis of digital media with respect to content understanding, indexing and creation of digital libraries. The *MultimediaMiner* [40, 41, 23] project and its predecessor projects [28, 24] at Simon Fraser University have constructed many image understanding, indexing and mining techniques in digital media. On related projects, there have been considerable advances in mining knowledge from spatial [20] and geographic [21] databases.

3. WORK IN PROGRESS

The objective of our *cinema miner* is to build a *framework* consisting of concepts, their implementation mechanisms, relevant algorithms and a test-bed to mine *interesting knowledge* contained in movies. We concentrate in mining higher level knowledge from streams using already available detection, identification and querying capabilities for underlying audio and video, with the expectation that our concepts, mechanisms and algorithms will be sufficiently generic and independent of the underlying media specific detection and identification mechanisms and thus will accommodate new advances in these areas.

Data contained in, or about a movie come from several sources. Firstly, there is some meta data, such as names of the crew including actors, director(s) etc, where the movie was filmed and other pertinent business data, criticisms, ratings, popularity measures, advertisements etc. These can be obtained in textual form. Secondly, the *story* may be available in some textual form too. Thirdly, there are audio and video tracks, which is the final outcome prepared for the audience. For our mining efforts, we assume that meta data and the movie is available, but plan to use the story in its textual form only as a means of validating our results.

The type of knowledge we are interested in mining from movies can be classified as follows:

Compositional Structure: This includes the number and sequencing of the movie into scenes, shots segments etc, and will be described shortly.

Interesting Events: Specific events that indicate emotions, mood, serenity, violence level etc., For example, crying may indicate sadness (or happiness), clapping or laughter may indicate happiness and bomb blasts or gun shots may indicate street violence.

Event Patterns: Such as *Violence is followed by sad scenes* with some probability.

Clustering Movies: Finding clusters of movies such as the one's with a *sad ending.*

Relationships Between Movies and Meta Data: These are relationships such as *Movies with sad endings produced by Studio X earns more than 20 million in their first year.*

We now explain each item in detail, and how our planned prototype is to mine them.

3.1 Compositional Structure

The structure of movies can be described using a cinematic hierarchy [10] such as scenes where each scene is divided in to a sequence of shots and each shot is divided into a sequence of frames, as given in Fig. 1. We use already developed techniques [39, 4, 26, 29, 31] to detect video and audio scene-cuts, and to develop a preliminary classification of movies as a collection of successive scenes. Further, the time stamp of the video track can be used to index into the audio track to obtain the corresponding audio transcript that *describes* the scene. By using commercially available audio to text translators [11], we can obtain the textual content of the spoken conversation. By analyzing the textual descriptions, we plan to construct cinematic hierarchies of movies.

Following audio-video technology can be used for the purposes of creating cinematic structure.

Detecting Scene Cuts: Using abrupt changes in chrominance, luminance and other visual parameters between successive video frames it is possible to detect scene cuts with a certain probability [39, 4].

Detecting Cinematic Artifacts: Using similar technology it is also possible to detect cinematic artifacts such as camera pans, tilts, zooming, etc within a scene [39, 31].

Detecting Shots within a Scene: By using cinematic artifacts occurring within a scene, it is possible to divide

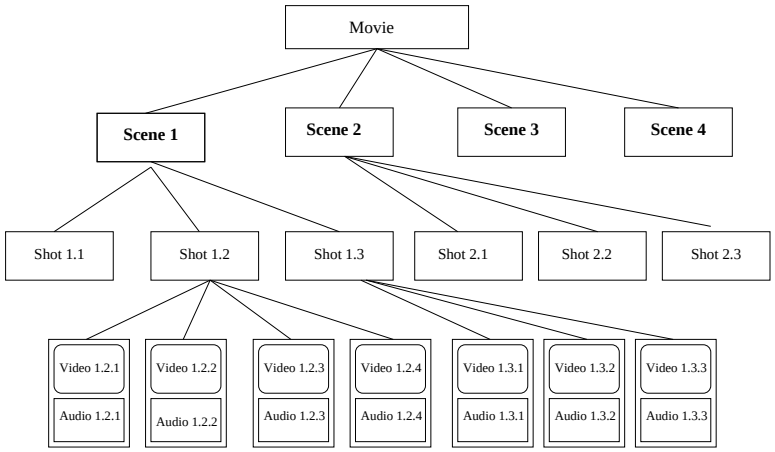

Figure 1: Movie Hierarchy

a scene into a sequence of *shots*. For example, one scene could be a *zooming in* action of 50 frames followed by a *zooming out* of 35 frames [31].

Similarly, there are techniques to segment the audio track consisting of four basic steps [31].

Separation of Music, Voice, Silence and Noise: By using frequency spectrum, relative loudness and other parameters of traditional sound analysis, it is possible to filter the sound track into music, voice. By using similar techniques, it is even possible to segment the music stream into different *beats*. These are currently available as either commercial products or research prototypes [35, 26, 29, 31].

Translating Voice into Text: There currently exist voice to text translators to translate the voice stream into the textual format. These are commercially available [11].

Understanding Text: By *understanding phrases* in texts [3], it is possible to categorize a text stream corresponding to a sequence of sentences by extracting a number of key words.

Recognizing Special Sounds: It has been shown that *gun shots*, *cries*, *clapping* and *bomb blasts* can be identified with some degree of certainty. Also there are research prototypes that can retrieve sounds that are *similar* to ones spoken to a microphone. (The analog of *query by image content*[13], called *query by humming*[14]).

We plan to *merge* knowledge gained by video segmentation, audio segmentation and textual understanding to characterize each movie consisting of a four level hierarchy as given in Fig. 1. At each level of the hierarchy above that of frames, we collect parameters, such as the sequence of camera artifacts, key phrases, special sounds (such as laughter, cries, gun shots, bomb blasts etc) and standard graphics parameters such as average RGB, luminance, chrominance etc. In addition, textual meta data about movies such as the director, rating, popularity, year of production, amount of money spent, studio used, country of production type of movie (i.e. thriller, children's movies, biography, romance, war, ethnic conflicts, Medieval story etc) actors, awards won etc are also available. By using these statistics we plan to mine for the following type of knowledge:

Frequent Episodes: What are the most frequent episodes in movies, or in a movie? Following standard terminology in temporal data mining, an episode consists of a sequence of *events*, where we consider an event to be any item in the movie hierarchy, such as frame, shot, scene or act. Notice that in order to apply known algorithms for mining for frequent episodes we need to enhance them to accommodate the following differences:

Basic Events: Identification of basic events has the following two added complexities.

Compound Nature: In the work of Manilla, Toivonen [25] and others, identification of basic events do not pose any problems: i.e. the associated textual predicate indicates if the event was present or not. In our case, the lowest level of events are complex objects consisting of at least one component of audio and video.

Probabilistic Nature: As stated in [25], the identification of basic objects is non probabilistic: i.e. either they were present or not. In our case, there is a non-perfect probability associated with them, as audio and video detection methods does not guarantee an absolute decision, but comes with an associated probability.

In order to address our mining requirements, we are re-addressing the basic formulation of *episode mining* to account for basic mine-able events consisting of components being identified with non-perfect probabilities. Our long term goal is to

extend this work to a case where one or more of these components are *missing* and the assumption of perfect temporal synchrony is no longer valid due to network based media delivery.

Hierarchy: There is a hierarchy of events such as frames, shots, scenes, acts that belong to different levels of granularity. Consequently, using patterns of events at a lower granularity being contained in larger episodes with coarse grain granularity need to be investigated. This has to be done using the knowledge gained at the lower level to help the mining at a higher level. This is similar to the issue faced in conventional episode mining, but requires some enhancements to account for probabilities associated with object identity.

Mining for Trends Within Movies: We plan to investigate trends such as *Do all movies begin happily and end sadly?* or vice versa. In mining for such knowledge we need to characterize *happy beginnings* or *sad endings*. In detecting such trends, we can use (for the lack of better word) a *happiness* index such as the mixture (or percentage mixtures of) laughter or crying within the scene, key words or phrases associated with the scene matching those that have been pre-characterized as indicating happiness, and the existence of special cinematic effects or average RGB colors, and beats in music). Notice that this is equivalent to classifying movies, according to certain criteria. E.g., a movie can be deemed *happy* or not; *violent* or *non-violent*. Once these concepts are associated with a series of parameters (as discussed before for *happiness* index), one can use classification techniques [22, 30, 5, 6, 18, 17] to achieve this task. Notice also that a movie can belong to more than one class: e.g., a movie can be *happy*, and *adult-oriented*. There are enough movies to verify the accuracy of potential predictions such as *A bomb blast is followed by two minutes of sad scenes*, or at a higher level *In action movies, an act of violence is followed by two acts more (perhaps retaliatory or punitive) of violence*, and the validity of such claims are to be tested by user studies.

Association Rules Mining: We plan to develop techniques to mine for association rules [1, 2, 15] at every level of the hierarchy. For instance, one can find that certain kinds of objects in a frame representative of a scene are often associated with other kinds of objects in the frame that represents the next scene. At the higher level, one can discover association rules among high-level concepts and features such as directors, amount of money spent, type of movie, *trend* (as mined by our algorithm) period, money earned, studio of production, and so on. Notice that applying standard association rules techniques to multimedia not only uncovers knowledge, but also helps in the object recognition problem, by potentially enhancing the probability assigned to the object. For instance, if one knows that objects type A and B occur frequently together (within a window of δ frames), and it is discovered in the movie that is being mined currently that an object, believed to be of type A and and object believed to be of type B are occurring frequently together, the probability of being of type A, attached to the first object and the probability of being of type B, attached to the second can be strengthened (via a heuristic procedure).

We also plan to extend our association rule mining for *quantitative knowledge*, which are customarily called quantitative association rule. An example would be *A bomb blast lasting T minutes is followed by two acts of mourning*, but *a gunshot is followed by crying in two consecutive scenes of the same act.*

Clustering Movie Trends and Categorization: We plan to cluster movie trends using our hierarchical information. Using the information obtained from mining trends and association rules, one can form a vector of features that describe the movie. Then, clustering algorithms [19] can be applied to group movies into similar classes.

4. PROTOTYPE CINEMA MINER

Fig. 2 shows the structure of our prototype. As stated in Section 1, it consists of using already available audio and video analysis techniques to separate and identify components and datum that are used in the mining process. The hierarchy constructor, text analyzer and movie miner are the components that will receive our concentrated attention in this project. Each of these units can be considered a module in a prototype we plan to build in stages as a result of this research agenda.

Textual Analysis of Audio: As shown in Fig. 2, this module is fed voice extracted from the audio stream with appropriate time stamps. It is then fed to text off-the-shelf text analysis module, where the audio will be translated to a parse tree of text. The parse tree can be used to analyze the *story* to some extent. We also propose to use selective phrases and words that are indicative of noteworthy or interesting events.

Cinema Miner: This module collects all the information in the form of a feature vector, including those detected by the video and audio analysis components, and mine for *knowledge* out of them. It is divided into trend prediction, event sequence mining, association rules and clustering sections, as described earlier.

User Validation of Mined Data: We plan to carry out user surveys to figure out the validity of mined knowledge from movies, in the spirit of perception studies related to multimedia usability [37].

5. CONCLUSIONS

We have presented a blueprint for mining information in and about movies that has become feasible by combining existing technology in image, audio and text analysis and understanding. As stated in the introduction, even under the assumptions of perfect time wise synchrony, mining common patterns existing in a collection of fused media streams present a challenge that goes even beyond the difficult problems of image, text and audio understanding. Even so the more important issues is to find out if common concepts used in mostly text based mining such as clustering, event patterns and commonly occurring pairs would be sufficient to describe common patterns and hidden knowledge available in cinema.

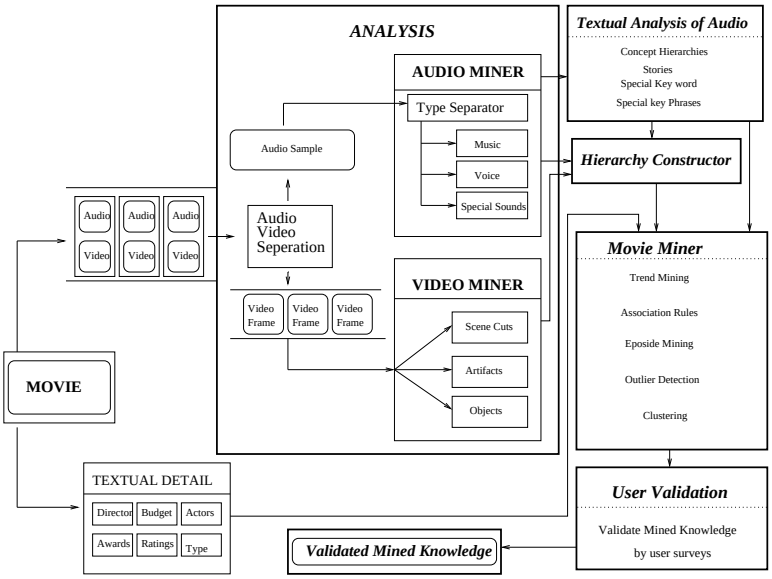

Figure 2: Prototype Cinema Miner

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A.Swami. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, may 1993.

[2] R. Agrawal, H. Mannila, R.Srikant, H. Toivonen, and A.Inkeri. Fast Discovery of Association Rules. In U.Fayyad, G. Shapiro, P. Smyth, and R. Uthurusamy, editors, *In Advances in Knowledge Discovery and Data Mining*. AAAI press, 1996.

[3] R. Beckwith, F. C., D. Gross, K. Miller, G. A. Miller, and R. Tengi. Five papers on wordnet: Special issue of the journal of lexicography. *Journal of Lexicography*, 3(4), 1990.

[4] J. S. Boreczky and L. A. Rowe. A comparison of video shot boundary detection techniques. *Journal of Electronic Imaging*, pages 122–128, September 1996.

[5] L. Breiman, J. Friedman, R. A. Olsen, and C. Stone. *Classification and Regression Trees*. Chapman and Hall, 1984.

[6] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998.

[7] M. Christel, S. Stevens, and H. Watlar. Informedia digital library. *Communications of the ACM*, 34(9):57–58, April 1994.

[8] The cuevideo project. http://www.almaden.ibm.com/cs/cuevideo/index.html.

[9] A. Czyzewski. Mining knowledge in noisy audio data. In *Proceedings of the Second International Conference of Knowledge Discovery and Data Mining (KDD'96)*, pages 220–225, Portland, Oregon, 1996. AAAI Press.

[10] G. Davenport, T. A. Smith, and N. Pincever. Cinematic primitives for multimedia. *IEEE Computer Graphics and Applications*, July 1991.

[11] Audiomine by dragon systems inc. Available at `http://dragonsys.com`.

[12] W. Effelsberg. Automatic movie content analysis. http://www.informatik.uni-mannheim.de/informatik/pi4/projects/MoCA, 1994.

[13] M. Flickner, H. Sawhney, and W. Niblack. Query by image and video content: The qbic system. *IEEE Computer*, 28:23–32, 9 1995.

[14] A. Ghais, J. Logan, D. Chamberline, and B. C. Smith. Query by humming: Musical information retrieval in an audio database. In *Proceedings of the Third ACM International Conference on Multimedia*, pages 231–236. ACM Press, 1995.

[15] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proceedings of the 1995 Intl. Conf. on Very Large Data Bases (VLDB'95)*, pages 420–431, September 1995.

[16] I. Haritaoglu, D. Harwood, and L. S. Davis. w^4: Who? when? where? what?: A real time system for detecting and tracking people. *FGR'98*, 1998. Available at `http://umiacs.umd.edu/users/lsd/vsam/Pubs.html`.

[17] T. Joachims. A probabilistic analysis of the rocchio algorithms with tfidf for text categorization. In *In Proceedings of the Intl. Conference on Machine Learning*, 1997.

[18] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In

Proceedings of the 10th European Conference on Machine Learning, 1998.

[19] D. Keim and A. Hinneburg. Clustering Techniques for Large Data Sets - From the Past to the Future. Tutorial session in ACM SIGKDD International Conference On Knowledge Discovery and Data Mining. San Diego, California, June 1999.

[20] K. Koperski, J. Adikary, and J. Han. Spatial data mining: Progress and challenges. In *SIGMOD Workshop on Research Issues on Data mining and Knowledge Discovery ((DMKD'96)*, pages 27–32, Montréal, Canada, 1996.

[21] K. Koperski, J. Han, and Adhikary. Mining knowledge in geographical data. *Communications of the ACM*, 1998.

[22] L. Kubat, I. Bratko, and R. Michalski. *Machine Learning and Data Mining, Methods and Applications.* John Wiley and Sons, 1998.

[23] Z.-N. Li, O. R. Zaiane, and Z. Tauber. Illumination invariance and object model in content-based image and video retrieval. *Journal of Visual Communication and Image Representation*, 1998.

[24] Z.-N. Li, O. R. Zaiane, and B. Yan. C-bird: Content-based image retrieval in digital libraries using chromaticity and recognition kernal. In *International Workshop on Storage and Retrieval Issues in Image and Multimedia Databases, in conjunction with the 9th International Conference on Database and Expert Systems (DEXA'98)*, Vienna, Austria, 1998.

[25] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovering generalized episodes using minimal occurrences. In *Proceedings of the second International Conference Knowledge Discovery and Data Mining*, pages 146–151. AAAI Press, 1996.

[26] K. Minami, A. Akutsu, and H. Hamada. Video handling with music and speech detection. *IEEE Multimedia*, pages 17–25, july-September 1998.

[27] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura. Video handling with music and speech detection. *IEEE Multimedia*, pages 17–25, July-September 1999.

[28] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the International Conference of Very Large Databases (VLDB'94)*, pages 144–155, Santiago, Chile, 1994.

[29] S. Pfeiffer, S. Fischer, and W. Effelsberg. Automatic audio content analysis. In *Proc of ACM Multimedia*, pages 21–30, ACM Press, New York, 1996.

[30] R. Quinlan. *C4.5 - Programs for Machine Learning.* Morgan Kauffman, 1993.

[31] C. Saraceno and R. Leonardi. Audio as a support to scene change detection and characterization of video sequences. In *Proceedings of the ICASSP*, IEEE Computer Society Press, 1997.

[32] S. Satoh, Y. Nakamura, and T. Kanade. Nameit: Naming and detecting faces in news media. *IEEE Multimedia*, pages 22–35, January-March 1999.

[33] S. Sirinivasan, D. Ponceleon, and D. Amir, A. Petkovic. What is that video anyway?: In search of better browsing. In *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, pages 388–392. IEEE, IEEE Press, June 1999.

[34] S. Srinivasan, D. Ponceleon, and D. Petkovic. Towards robust features for classifying audio in the cuevideo system. In *ACM Multimedia 99.* ACM, ACM Press, November 1999.

[35] Y. Tonomura, A. Akutsu, Y. Taniguchi, and G. Suzuki. Structured video computing. *IEEE Multimedia*, pages 34–43, Fall 1994.

[36] D. Wijesekera and D. Barbara. *Mining Multimedia Datasets*, chapter F7, Handbook of Data Mining. Oxford University Press, 2000.

[37] D. Wijesekera, J. Srivastava, A. Nerode, and M. Foresti. Experimental evaluation of loss perception on continuous media. *Multimedia Systems*, 7:486–499, October 1999.

[38] E. Wold, T. Blum, and J. Wheaton. Content-based classification, search and retrieval of audio. *IEEE Computer*, 28:27–36, 9 1996.

[39] R. Zabin, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying production effects. *ACM Multimedia Systems*, 7:119–128, March 1999.

[40] O. R. Zaiane, J. Han, Z.-N. Li, S. H. Chee, and C. J. Y. Multimediaminer: A system prototype for multimedia data mining. In *Proceedings of the 1998 ACM-SIGMOD Conference on Management Data (system Demo)*, volume 38, Seattle, Washington, 1998.

[41] O. R. Zaiane, J. Han, Z.-N. Li, and J. Hou. Mining multimedia data. In *Proceedings of the CASCON'98: Meeting of Minds*, pages 27–32, Toronto, Canada, 1998.