REAL TIME MONITORING OF RADIO AND TV BROADCASTS

Werner Haas, Harald Mayer and Georg Thallinger

JOANNEUM RESEARCH
Institute of Information Systems & Information Management, Graz, Austria
{Werner.Haas, Harald.Mayer, Georg.Thallinger}@joanneum.at

ABSTRACT

This paper describes work being done in two European projects. While content based indexing and retrieval methods have initially mostly been applied to the audio and video archive area, both projects target the media monitoring market. This is motivated by the size, demand and structure of this market. On one hand, the yearly worldwide revenues are remarkably high, and on the other hand, also methods that are maybe not yet perfect for the absolute requirements of very large archives may be well suited to enhance or replace existing manual solutions in media monitoring.

The project RAA provides an example for the recognition and monitoring of audio clips, based on digital fingerprinting technologies. In the project DETECT real time recognition of advertisements and company logos for producing comparative brand statistics is being developed. Approach and results, specific technical problems for this application area as well as applicability and market relevance of the developed solutions are discussed.

1. INTRODUCTION

Content-based multimedia indexing and retrieval has been around for some time already. A lot of research and development has been carried out on the core topics of analysing content, extracting features and basing elaborate query models on those features later on. Two general questions have only recently been addressed by a broader research community. The first one is, whether the conventional search and query approach, brought in from the conventional and highly developed area of text analysis, search and query, is really the only and an efficient one for multimedia. The second question is, for which market existing technologies can already be applied.

The project work described in this paper mainly addresses the second question. Of course – and similar to the search and query approach - from the beginning researchers have concentrated onto the obvious and basic need of multimedia archives. During the 1990s quite a number of companies offering video archive solutions have come up and disappeared again while only a very few have survived in this market.

Media - and among them again the audiovisual media – are enjoying an ever-increasing important role and influence. Starting from the theoretical point that media just report in a neutral way what is going on in the world and that they reflect the public opinion, they have since their beginning developed into a very efficient means to influence public opinion, consumer behaviour and our social environment in general. Not only is the public opinion influenced by e.g. election campaigns, and not only do advertisements try to modify our buying habits. Also in a very general sense, our social behaviour, usage of certain words and their pronounciation and spelling, our cultural preferences and also our style of communication are being influenced by media.

Many of these social changes are influenced by the media in a not easily and directly measurable way, e.g. in the simplest case by repeatedly reporting about a topic in a positive or negative manner. Knowledge management and knowledge extraction technologies are already being applied to such problems. Issues of direct presentation of logos and advertisements to the watcher and/or listener can be more easily tackled by current CBIR technologies. In particular, there is great interest of the media monitoring industry to know as exact as possible at which time certain content (advertisements, logos, jingles etc.) was broadcast in which channel. This support to identify content or parts of it has been implemented in the two projects to be described in this paper.

What are now typical applications in the media monitoring area? Among them are analysis and recognition of:

- advertisements,
- songs,
- market information, news, PR.

For broadcasters, also monitoring of:

- compliance and
- quality of service

are of importance. For government agencies, financial institutions, insurancies and large industrial groups

- news monitoring,
- business information monitoring,
- trend recognition and monitoring and
- (real time) decision support

are the most valuable tools expected to be provided by CBIR methods and software.

As an illustration of that, a few data about the market are given in the following. In 2001 in Germany alone approximately 2.4 million TV spots were broadcast. There were about 13.000 different spots, of which 40 % were produced in this year (Source: Ogilvy Group Germany). Yearly revenues for the German advertisement market in 2002 were around 8.200 M€ just for TV and radio. Approximately the same amount was spent for classical paper advertising (Source: Nielsen Media Research GesmbH). Worldwide, for Internet advertising approximately 13.000 M€ were spent, at a much higher growth rate however, predicting around 35.000 M€ for 2004 (Source: EmediaPlan).

2. IDENTIFYING MUSIC TITLES

Due to the ever increasing amount of distribution channels for music content, it becomes more and more important for content owners (i.e. producers, authors) to know where and how often their works is broadcasted or made available over Web channels. In former days content owners, represented by national collecting societies, relied on proper usage reporting of the few national broadcasters and making some random samples. Nowadays monitoring is a time consuming task and there is a real need to automate this process.

In this context the European research project RAA (Recognition and Analysis of Audio) was established to develop a complete system solution to identify music titles as they are broadcasted. The main requirements on this system was robustness against typical transmission effects and that in contrast to watermarking approaches the music titles are not marked with any embedded information.

In the following sections we describe the manifold requirements to a recognition system, the fundamentals of our approach and the components allowing to provide tailored monitoring products for the market.

2.1. Audio fingerprinting

Approaches fulfilling the above requirements are falling in the category of content-based identification (CBID) and are commonly known as audio fingerprinting. In a general fingerprinting scheme, the system generates a unique fingerprint of the audio material based on an analysis of the acoustic properties of the audio itself. Several features can be found in literature for the characterization of audio: energy, loudness, spectral centroid, zero crossing rate, pitch, harmonicity, spectral flatness and Mel-Frequency Cepstral Coefficients (MFCC's).

Fingerprinting systems do not rely on embedded information within the music signal or on any other external metadata information transmitted next to the music itself (e.g. like it is possible using RDS – Radio Data System for VHF/FM broadcasting or within DAB – Digital Audio Broadcasting).

The fingerprint developed within RAA is called AudioDNA, as it is similar to biological DNA sequences. One can imagine the AudioDNA as a stream of letters, where each letter additionally has also a duration attached. This AudioDNA is much smaller in size compared to the original signal, e.g. approximately 10 Kbytes per minute. Identifying titles is performed by comparing the fingerprint of an observed signal against a database of reference fingerprints stored in database of previously acquired music titles.

2.1.1. Related work

There are several approaches for content-based identification and search of audio material. E. Wold et al. [1] from Muscle Fish describe a system for finding similar sounds to a given example. This system extracts time-varying properties from sampled sound files and for each property the mean, variance and autocorrelation over the entire file is recorded. At the time of their publication the system was used for comparison of noises, like scratches, bells and laughing but was not used for whole song identification.

In the meantime, during our project development phase, several research organisations and companies are providing music identification technology and services, like AudioID from Fraunhofer [2] or a system developed by Philips [3, 4]. In addition there are other comparable approaches like tuneprint [5], eTantrum [6], cantametrix [7] and audible magic [8]. However, due to the commercial orientation there are not many details available about the methods and technologies used by the last mentioned approaches.

2.2. Technical objectives

Several requirements had to be fulfilled by the developed identification system. First and foremost are the robustness requirements on the identification itself. Having two fully identical audio signals the comparison obviously is not too difficult. However, in practise there are several different aspects which have to be taken into account. Depending on the audio source (e.g. CD, Radio, Internet audio sources, etc.) the actual audio signal may be very different from a master source. The main reasons for such differences are:

- Audio storage: Signal-to-noise ratio, digitization parameters and frequency range of the storage media, influence the degeneration of the audio signal. Especially for audio sources from the Internet, differences are caused by the various compression formats and compression rates which are used for saving bandwidth.
- Transmission effects: The broadcasting techniques (e.g. frequency modulation (FM), amplitude modulation (AM)) specify amongst other things the bandwidth and the signal-to-noise ratio of the observed audio signal. The Internet compression technologies used within streaming media protocols could also be assigned to this category.
- Manipulative effects: Radio stations use complex sound processing to apply their own sound characteristics. For example they often enlarge the stereo base, which enhances the impression of stereo listening, or they play the songs faster (pitching).

In addition to these general types of effects, the developed fingerprint method must also work if only parts of a music title are available/broadcast. Broadcasting stations very often shorten the title or start a title after some seconds of the title.

Also cross-fading between titles and speakers talking over a music title are a real challenge for such identification systems.

In terms of performance the system must be able to do identification on several broadcasting streams in parallel. Our goal was to allow the observation of 50-100 channels in parallel against a reference database of at least 100.000 titles. Additionally the system should be highly scaleable and based on standard PC technologies, without any special (custom) hardware needed for processing the audio signal.

2.3. RAA architecture

The developed system architecture is shown in Figure 1 below. The most important parts are the AudioDNA extractor, which continuously calculates the fingerprint of a digitised audio signal, and the AudioDNA matcher, which compares incoming AudioDNA streams against references fingerprints stored in a database.

Based on these main functionalities the system components are implemented. The AudioDNA extraction procedure is mainly based on statistical pattern matching algorithms and the procedure is based on speech recognition technologies. More details on the algorithms can be found in [9].

An approximate matching algorithm had to be developed as the reference fingerprint will always be different to the observed one due to the broadcasting effects mentioned earlier. As such only short substrings of the observed signal may be identical to the reference signal, however the distance between identical substrings has to be similar in both fingerprints if they represent identical titles. An additional issue is that no reasonable metric between audio genes can be established and used in the matching process.

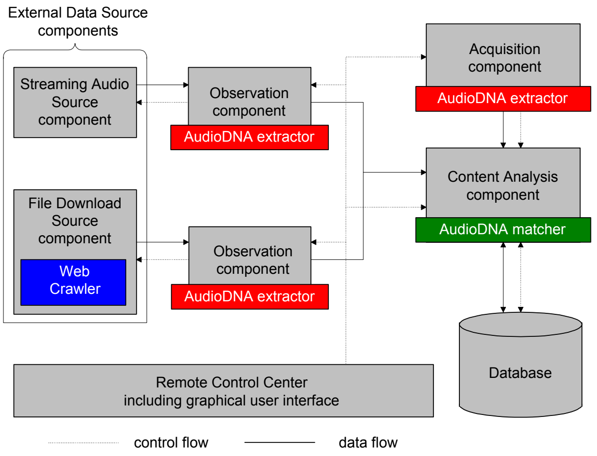

External Data Source components
Streaming Audio Source component
File Download Source component
Web Crawler
Observation component
AudioDNA extractor
Observation component
AudioDNA extractor
Acquisition component
AudioDNA extractor
Content Analysis component
AudioDNA matcher
Database
Remote Control Center including graphical user interface
control flow
data flow

Figure 1: System architecture.

However, this is a characteristic problem in text searching applications and in the field of computational biology. The matching algorithm applied consists of two main processing steps, which use the same principle method as FASTA [10], one of the most effective practical database search methods for biological gene sequences:

1. First the search space in the database is reduced by exact matching of short sub sequences.
2. On the positions of the exact matches the similarity of a longer sequence is examined by an approximate matching algorithm.

In contrast to text strings or biological gene sequences, the genes of our AudioDNA additionally contain time information. Each audio gene represented by a letter corresponds to a specific time period of the acoustic signal (see Figure 2)

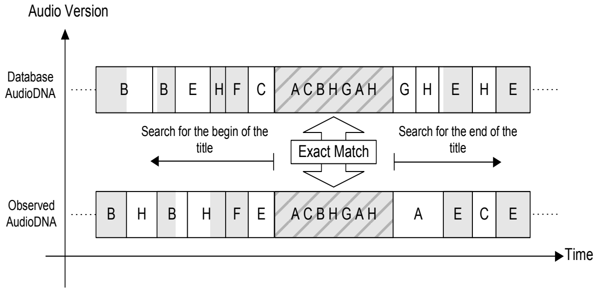

Audio Version
Database AudioDNA
B B E H F C ACBHGAH G H E H E
Search for the begin of the title
Exact Match
Search for the end of the title
Observed AudioDNA
B H B H F E ACBHGAH A E C E
Time

Figure 2: The exact matching result is used to detect equal genes which have a time overlap (indicated by the grey areas).

To detect similarities of longer sequences at the position of the exact matches, an approximate matching algorithm is used. Starting from the exact matching position the audio genes are consecutively compared in each direction (see Figure 2). If the similarity of the sequence similarity falls short of a predefined threshold in a given time period, the start or the end of the matching sequence and therefore of the audio title is detected.

The calculation of the similarity of sequences is a standard application of dynamic programming [11]. The time information of the audio genes is used to optimise the approximate matching algorithm. More details can be found in [12].

2.4. Performance results

The developed identification system satisfies the performance objectives laid out in the beginning, thus a single analysis station can do parallel comparisons of 50 input channels against at least 50.000 reference titles. A single observation station is able to observe at least 4 audio channels in parallel. The observation stations are connected to the analysis station using standard local network connection. And as the system architecture is scaleable, more than one analysis station can be utilised.

The robustness of the RAA system was heavily evaluated during the project, using a reference database of some 50.000 titles. The categories of these reference titles were 3% blues, 5% classic, 20% pop, 15% rock, 5% hard rock, 5% jazz, 15% techno/dance, 30% film score and 2% country/western music. In addition to manually processed test material also several popular radio stations in Germany and Netherlands and MTV television have been observed. The evaluation led to the following results. The system is robust against:

- **Audio coding / compression:** MPEG-1 audio layer 3 compression (MP3) down to 48 kbit/s, RealAudio (RM) from Realnetworks down to 48 kbit/s, Microsoft Windows Media Audio (WMA) down to 48 kbit/s
- **Drop-outs / missing sections:** Missing sections within an observed music title do not prevent identification, provided that the sections before and after the missing section can be identified independently
- **Mixing / crossfadings:** The same holds for mixing to titles or crossfadings between titles. The sections before and after such an effect will be identified correctly
- **Pitching and other broadcast effects:** The AudioDNA used in RAA is tolerant against

pitching (up to 6%) and other effects commonly used in radio broadcast. In combination with the ”inexact” database comparison the result is even more tolerant against such effects

- **Sections:** RAA is able to recognise single sections (parts) of a title, which was brought into the reference database in its full length. However, reliable identification requires a sufficiently long and characteristic part of the title. Typically the identification rate using characteristic parts with a duration of 10 seconds will be >99,5 %. The amount of the most troublesome ‘false positives’, i.e. wrong identification, is limited to a maximum of 0,1%.

2.5. Available products

Based on the feedback of potential market-players several key-turn product solutions have been compiled.

Amongst them is the “Commercial Monitor”, providing an efficient way for detection of commercials in broadcast audio-signals. This can be used either for TV or radio. Commercials in the sense of the product are audio-clips with a duration between 5 and 60 seconds. The system consists of 3 PC, allowing parallel observation of 3 input channels and a reference database of 10.000 commercials.

Next solution is the “Music Monitor”, which is similar to the above solution, but allowing more titles in the reference database.

This basic configuration can be extended with both the “Monitor Extension”, allowing the observation of more input channels in parallel, and the “Analysis Extension”, allowing more titles to be stored in the reference database.

In addition to these turn-key solutions also some single components are available, allowing system integrators to include RAA functionality within their environment.

2.6. Acknowledgments

The main work described in this section was done within the European project RAA, which was partially funded by the European Commission (IST-1999-12585) and was also supported by the Austrian Federal Ministry of Science and Research. More information on the project can be found at http://raa.joanneum.at and http://www.musicreporter.net.

The authors would like to thank all other partners of the consortium for their cooperation, especially Eloi Batlle and Pedro Cano from Pompeu Fabra University, Barcelona and Walter Plaschzug from HS-art digital service, Graz.

3. BRAND MONITORING

Statistics on brand visibility – how often a logo has been shown e.g. during a sports event - are of imminent importance for the advertising industry. Based on such statistics the effectiveness of advertisement campaigns could be measured or the value of a logo placed on a publicity board estimated.

The vast amount of video material produced and broadcast every day require new approaches to compile statistics about the content. Due to the large amount and costs human beings are not able any more to fulfil this task. Only approaches based on algorithms running autonomously are able to gather such statistics. The principle goal of the EU-IST cognitive vision project DETECT is to implement a general platform for such autonomous algorithms and to provide concrete test applications, such as semantic block detection, brand detection and agent tracking & recognition. The challenge on image analysis is to perform object detection in arbitrary outdoors scenes disregarding illumination changes or weather conditions, and continuously maintaining a certain quality of service (e.g., accuracy, robustness, speed). This demands for the development of cognitive vision methods in order to apply cascaded object detection and recognition based on receptive field methodology, and to focus attention from scene and object related context in order to structure the detection process.

3.1. DETECT system architecture

The overall system consists of 4 parts. The **detection engine** is the core component within DETECT's system architecture and is divided into two major parts, the framework and a number of modules executed by the framework comprising an algorithm. The modules and their interconnections form a data flow graph. The modules are executed asynchronously; computation results are buffered temporarily by the detection engine. Every module performs a specific computational task (e.g. histogram calculation); together the modules solve a high-level task (e.g. logo detection). Results computed in the detection engine are saved in the **database**. The **observation and control interface** (OCI) serves as a front-end for the detection engine. The OCI may run on a separate or the same machine as the detection engine. It displays current computation results and allows control by the operator of the ongoing computation task performed on the detection engine. The **statistics engine** is the interface to end users. Via this interface the end user may retrieve compiled statistics from the database (e.g. how often a specific logo was viewable).

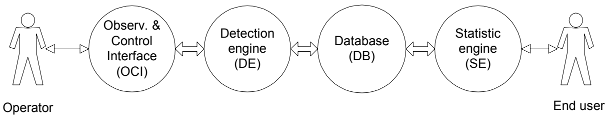


Figure. 3: Basic DETECT system architecture.

3.2. Cognitive vision issues

In computer vision, we face the highly challenging task to perform detection of relevant object events in outdoor environments, such as in sport broadcasts. Changing illumination, different weather conditions, and noise in the imaging process are the most important issues that require a truly robust detection system.

This project considers prediction schemes that would significantly improve the quality of service in real-time interpretation of image sequences. Research on video analysis has recently been focusing on object based interpretation, e.g., to refine semantic interpretation for the precise indexing and sparse representation of immense amounts of image data. Object detection in real-time, such as for video annotating and interactive television, imposes increased challenges on resource management to maintain sufficient quality of service, and requires careful design of the system architecture. Recent work on real-time interpretation therefore applies methodologies on the basis of an image-to-hypothesis mapping. In DETECT, we apply attention based mechanisms and a cascaded system framework to coarsely analyze the complete video frame in a first step, reject irrelevant hypotheses, and iteratively apply increasingly complex classifiers (e.g. as presented in [13]) with appropriate level of detail. In addition, context priming makes sense out of globally defined environmental features to set priors on object detection observable variables.

Investigations on the binding between scene recognition and object localization made in experimental psychology have produced clear evidence that highly local features play an important role to facilitate detection from predictive schemes. In particular, the visual system infers knowledge about stimuli occurring in certain locations leading to expectancies regarding the most probable target in the different locations. DETECT uses contextual cueing for spatial attention in object detection inspired by the human cognition model. The knowledge about forthcoming detection events has been built up in repeated processing on the scene before it is used to predict object occurrences. It thereby imposes significant performance improvements due to the reduced object search space.

A highly challenging issue is the integration of the individual visual processing components into a cascaded object detection system that is capable of working in time and at video rate, and maintaining the quality of service across appearance variations due to environmental conditions.

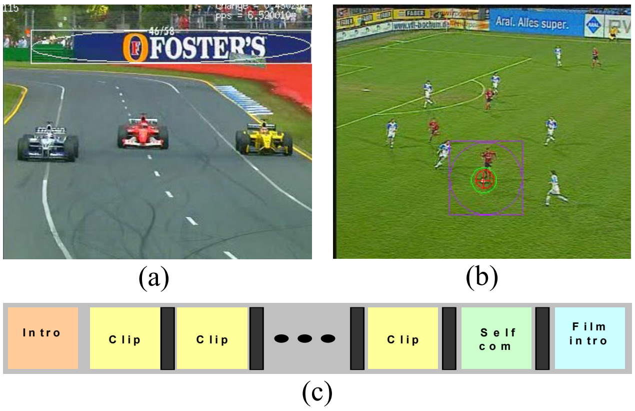
3.3. Applications

DETECT aims at three major application goals: detection of semantic blocks (commercial and movie categories), brand detection (gathering statistics of logo occurrences) and agent tracking & recognition (e.g. tracking of the ball, the players in sport games). All these concrete applications are based on the software components explained above.

Semantic block detection Semantic block detection focuses especially on detection of commercials. A typical structure of a commercial block is shown in Figure 4 (c). Previous work has focused mainly on temporal features like series of monochrome frames, shot ratio, motion patterns (e.g. [14], [15]). In addition to that the use of image content e.g. fade out of broadcasting service logo, text and logo appearance, for the use of commercial detection will be investigated.

Brand detection Advertisements and commercials are the main source of income for TV-stations all around the world. Detailed data about product placements and visibility of advertisements help the advertising-industry to find strategy for optimal product-placement and TV-broadcasters to collect advertising-fees more transparently then ever before (Figure 4 (a)).

Agent tracking & recognition Individually moving objects like persons, balls and cars once detected are automatically tracked in order to generate information about their position and size. This information can be used for automatic creation of inserts as well as basis for sports statistics. Proper definition of agents applied on a moving image sequence provides a way for automatic identification and recognition. Thus whenever tracking fails, there is a chance to resume after automatic recognition. Samples for such agents are sports-teams, race-cars and balls (Figure 4 (b)).


(a) (b)

(c)

Figure 4: (a) Detection of a logo object in a Formula One video stream. (b) Tracking of an agent (ball) in soccer (c) Typical structure of a commercial block.

3.4. Acknowledgments

The main work described in this section was done within the European project DETECT, which is partially funded by the European Commission (IST-2001-32157) and is also supported by the Austrian Federal Ministry of Science and Research. More information on the project can be found at http://www.detect-tv.com.

The authors would like to thank all other partners of the consortium for their cooperation, especially James L. Crowley from INRIA Rhônes-Alpes, Montbonnot, Lucas Paletta from the Institute of Digital Image Processing, JOANNEUM RESEARCH, Graz and Walter Plaschzug from HS-art digital service, Graz.

4. REFERENCES

[1] Wold, E., Blum, T., Keislar, D., and Wheaton, J. “Content-Based Classification, Search, and Retrieval of Audio”. IEEE Multimedia, Vol. 3., No.3, 1996, 27-36.

[2] Allamanche E., Herre J., Hellmuth O., Bernhard Fröbach B. and Cremer M., “AudioID: Towards Content-Based Identification of Audio Material”, 110th AES Convention, Amsterdam, The Netherlands, May 2001.

[3] Haitsma J., Kalker T., “A Highly Robust Audio Fingerprinting System”, ISMIR 2002, 3rd International Conference on Music Information Retrieval, Paris, France, October, 2002

[4] Kalker T., "Applications and Challenges for Audio Fingerprinting", presentation at the 111th AES Convention, NY, in the "Watermarking versus Fingerprinting" workshop, December 3, 2001.

[5] http://www.tuneprint.com

[6] http://www.etrantrum.com

[7] http://www.cantametrix.com

[8] http://www.audiblemagic.com

[9] Cano P., Batlle E., Mayer H., Neuschmied H., “Robust Sound Modelling for Song Detection in Broadcast Audio”, 112th AES Convention, Munich, Germany, May 2002.

[10] Perason, W. R. and Lipman, D. J. Improved tools for biological sequence comparision. Proc. Natl. Academy Science, Vol. 85, 1988, 2444-2448.

[11] Guisfield, D. Algorithms on Strings, Trees, and Sequences. Computer Science and Computional Biology (Cambrigde, 1999), Cambridge University Press, ISBN 0-521-58519-8.

[12] Neuschmied H., Mayer H. and Battle E., “Identification of Audio Titles on the Internet”, Proceedings of International Conference on Web Delivering of Music 2001, Florence, Italy, November 2001.

[13] Pelisson, F., Hall, D. and Crowley J.L.: “Brand Identification Using Gaussian Derivative Histograms”, Proceedings of International Conference on Computer Vision Systems 2003, Graz, Austria, April 2003.

[14] Lienhart, R., Kuhmunch, C. & Effelsberg, W., On the Detection & Recognition of Television Commercials, Proc. IEEE Conf. on Multimedia Computing and Systems, pp. 509 - 516, Ottawa, Canada, 1996.

[15] Sadlier D, Marlow S, O'Connor N. and Murphy N., Automatic TV Advertisement Detection from MPEG Bitstream. Journal of the Pattern Recognition Society, Vol.35, No.12 , December 2002.