

2.3.4 Content Based Identification (Fingerprinting)

Jürgen Herre²²⁴

I Introduction

Ever since the broad availability of efficient source coding methods (data reduction) and digital distribution channels (including the Internet), consumers have seamless access to an enormous amount of multimedia data. This includes audio material and still and moving pictures within a wide range of quality, ranging from “pre-view” (e.g. Internet radio) to broadcast quality. As a result, efficient handling of this considerable amount of data has become a challenge of its own (e.g. “how can I find desired material efficiently?”). This has led to the definition of a number of so-called *metadata* standards. Examples for such specifications include the Dublin Core initiative²²⁵, the SMPTE/EBU Dynamic Metadata Dictionary²²⁶, the P/Meta project of the European Broadcasting Union (EBU)²²⁷ and, more recently, the MPEG-7 standard²²⁸. The general idea behind these standards is to define data formats which provide a comprehensive description of the actual multimedia content in an interoperable way. Such meta-data (i.e. “data about data”) structures may include a wide range of descriptions of the origin and identity of the content, its structure, usage rules, and various perceptual or semantic aspects.

Among the many conceivable ways of characterizing a piece of audiovisual content, the unique description of the content identity based on its signal representation (so-called “content-based identification”) is of great importance. This functionality is frequently also referred to as *fingerprinting*²²⁹ and enables automatic identification (including title, author and other description of the works) of content which has been registered previously in an internal database of reference data. The topic of fingerprinting has received much attention recently in both research and commercial deployment and current technological development has shown that, depending on the underlying technology, reliable and efficient identification can still be achieved even for distorted input signals and large databases of multimedia material.

This article discusses the concept of content-based identification and the underlying technological challenges as well as some of its many attractive applications

²²⁴ Fraunhofer Institut für Integrierte Schaltungen, Erlangen, Germany.

²²⁵ Web site of the Dublin Core Metadata Initiative: <http://dublincore.org/>

²²⁶ See: SMPTE (2001).

²²⁷ See: Hopper (2000).

²²⁸ See: MPEG-7 Introduction (2001).

²²⁹ As a note to the reader it should be mentioned that the term *fingerprinting* is occasionally also used in the literature in the context of digital watermarking where the idea is to enable unambiguous identification of the content by imprinting a unique mark into the signal (rather than deriving a fingerprint from it). Unfortunately, this use of terminology may lead to considerable confusion and is, therefore, not endorsed by the author.

in the multimedia area. Owing to the underlying idea, the fingerprinting approach is very different in its nature from (and in fact in a sense complementary to) the concept of watermarking. Thus, the article is concluded by contrasting both approaches with respect to their use cases.

II The Concept of Fingerprinting

During the recent years, a number of technologies for fingerprinting of multimedia material were developed. In contrast to the identification of content based on embedded digital watermarks, fingerprinting is a “non-invasive” approach which does not require any modification of the original multimedia signal. The underlying idea consists of identifying the audio/image/video content directly by examining the characteristics of its signal representation using a *pattern recognition* process. As usual within the framework of pattern recognition, a *training phase* is required so that the characteristics of the items to be recognised are introduced into the system. This leads to a two-stage process (see Figure 1):

- During the *training phase*, characteristic features are extracted from a set of known reference items such that the extracted feature data forms a unique combination which allows for the unambiguous distinction of a particular item from all other entries. Such feature representations can be made extremely compact (e.g. several orders of magnitude smaller than MP3-compressed audio) and are frequently called *fingerprints*, *signatures*²³⁰ or *robust hashes*²³¹. For each item which should be recognised later by the system, such a fingerprint is generated and stored in a reference database together with some of its descriptive metadata. This metadata may be just enough for the identification of the individual item in terms of bibliographic reference (e.g. title name, artist) or may contain richer descriptions of the content.
- During the *recognition phase*, the signal to be identified (*query* item) is presented to the system and used for the extraction of a fingerprint in a way similar to that of the training phase. The actual recognition process is based on comparing this query fingerprint with the fingerprints that are stored in the reference database. The most “similar” fingerprint found in the database corresponds to the best matching (and most likely) reference item. As a result of this comparison process, the system delivers an indication of whether the presented signal has been successfully identified and, if this is the case, the database ID of the identified item together with a measure of the achieved recognition confidence. Furthermore, the metadata associated with this database item may be returned by the system.

²³⁰ See: Hellmuth, Allamanche, Herre, Kastner, Cremer, Hirsch (2001).

²³¹ See: Haitisma, Kalker, Oostveen (2001).

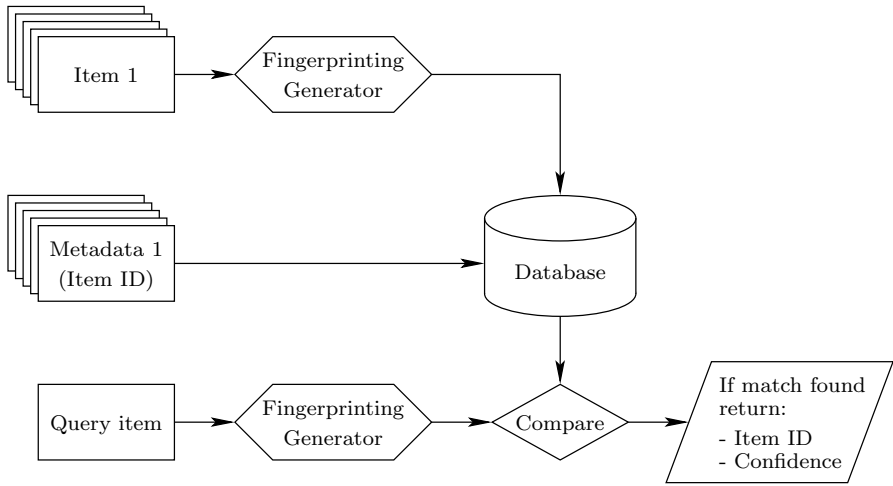


Fig. 1. Generic Structure of a Fingerprinting-based System for Content Identification

A good fingerprinting system should satisfy a number of requirements. The most important of them are briefly discussed here:

- Robustness:** When content-based recognition schemes are used in real world application scenarios, it is essential that correct identification is maintained also in response to a distorted query signal which is comparable to the requirement of robustness in the watermarking context. As an example, audio signals may have undergone simple modifications (such as level change, linear filtering/equalisation, bandwidth limitation, noise addition) or more complex alterations (such as MP3 data compression, GSM speech coding, watermarking, pitch and speed change). Furthermore, identification should also be possible if only arbitrary parts of the full signal are presented to the recognition engine (e.g. “15 seconds of audio starting at 1 minute 20 seconds into the song”) in a way analogous with the human ability of recognising excerpts. Finally, the system is required to reliably reject unknown content rather than confusing it with any of the registered reference items.
- Signature compactness:** Considering the fact that some applications of fingerprinting require recognition ability for millions of content items and that the system has to handle the corresponding amount of signature information, it becomes clear that the *compactness* of the signatures is of great importance. After the feature extraction process, only an extremely small fraction of the original information is retained. The signature data rates are usually several orders of magnitude lower than the rates used for audiovisual source compression since reconstruction of a representation in the original audiovisual domain is not intended. Nonetheless, this fraction of data has to support a reliable distinction between the query item and all other items in the reference database.

- *Search Speed*: As an additional requirement, the search process must be completed within a time period that is acceptable in the context of the application even for large search spaces (i.e. large signature databases).
- *Speed of signature extraction*: For the creation of comprehensive signature databases (e.g. several million items) from available audiovisual material, a fast signature extraction process is of major importance in order to complete the task within a realistic time frame. As an example, for many audio fingerprinting applications an extraction time that is significantly below the playing time of the item may be desired.

As is known to be true for many technologies, not all desired performance parameters of a system can be optimised independently. In the field of fingerprinting, a dependency between the recognition robustness and the compactness of the signatures can be observed²³². Usually, high robustness with respect to signal distortions can be achieved only by accepting an increased signature datarate. Conversely, applications which deal only with slightly distorted signals permit usage of extremely compact signature formats. It is instructive to compare this two-way trade-off to the three-way dependency between robustness, datarate and perceptual transparency that has been formulated in the watermarking context²³³.

From the principles discussed so far it becomes clear that fingerprinting-based systems achieve an identification of query signals based on their *similarity* with respect to the items contained in a reference database (*similarity-based* matching). It is mainly the type of features employed for signature extraction which determines what is interpreted as similarity by the system and thus plays a crucial role in the performance of the system. Depending on the nature of input signals, a large number of different features have been used for this purpose, such as color/shape/motion for video input, and spectral/temporal characteristics for audio signals. As fingerprinting systems are typically optimised for the best possible distinction between several items and reliable recognition even in the context of signal distortions, items that appear similar to a human observer (in whatever sense) are typically classified as “dissimilar” by such systems. Examples from the area of audio fingerprinting are:

- A “known” music title which is performed by a different artist,
- A person singing or humming the title’s melody or
- A different performance (e.g. “live version”) even by the original artist.

Thus, it is important to understand that such systems are generally *not* optimised to model subjective similarity of content, but to achieve a reliable distinction between different content even though it may be considered similar by humans. This approach is in fact a necessary requirement for enabling identification of different artistic instantiations of the same theme.

²³² See: Kastner, Allamanche, Herre, Cremer, Grossmann, Hellmuth (2002).

²³³ See: Neubauer, Herre (2000).

III An Example: MPEG-7 Audio Fingerprinting

The increasing interest recently in content-based identification has also stimulated a number of publications and systems in the area of audio fingerprinting which try to establish a presence in the market place, e.g.²³⁴. For the vast majority of these systems a comprehensive description of the underlying technological foundations is not available in view of their commercial background. Therefore, this section will illustrate the concept of audio fingerprinting by examining a current development based on the recent MPEG-7 Audio standard²³⁵. Owing to the standard-based approach, this technology relies on a fully specified open format of the signature data²³⁶.

- *Signature features*: The system relies on the so-called *spectral flatness* (SFM)²³⁷ of the audio signal which is calculated within subbands of 1/4 octave width each. Roughly speaking, this feature relates to the presence of tonal components within the subbands. The signature contains SFM data for a user selectable number (at least 4) of subbands starting at a frequency of 250Hz.
- *Computational complexity*: An extremely fast signature extraction process can be achieved by using current Personal Computers (ca. 100 times faster than the actual playing time of the music title). The recognition part can be implemented efficiently on both general purpose computing platforms (PCs) as well as portable devices, such as Personal Digital Assistants (PDAs).
- *Recognition performance*: The system is robust to common signal distortions and is able to identify arbitrary excerpts of the query item independent of its temporal offset. For a test database of 85.000 music titles and a number of signal distortions, a recognition performance of typically better than 99.7% was achieved.
- *Scalability*: Different fingerprinting application scenarios are usually associated with different robustness requirements and, thus, different optimal “operating points” in the trade-off between the compactness of the signature and its robustness. Using MPEG-7 Audio signatures, this can be accounted for by *scaling* several parameters of the signature in response to the application requirements and thus controlling the recognition strength (and, thereby also, the data rate). In this way, a range of signature data rates between 2

²³⁴ See: Auditude web site: <http://www.auditude.com>; Wold, Blum, Keislar, Wheaton (1996); etantrum music id. web site: <http://www.etantrum.com>; Haitisma, Kalker, Oostveen (2001); Kurth, Clausen (2001); Neuschmied, Mayer, Battle (2001); Moodlogic Inc. web site: <http://www.moodlogic.com>; Relatable web site: <http://www.relatable.com>; Shazam Entertainment Ltd. web site: <http://www.shazam.tv>; Tuneprint (robust psychoacoustic fingerprinting) web site: <http://www.tuneprint.com>.

²³⁵ See: MPEG-7 (2001); Lindsay, Herre (2001).

²³⁶ See: Allamanche, Herre, Hellmuth, Fröba, Cremer (2001); Hellmuth, Allamanche, Herre, Kastner, Cremer, Hirsch (2001); Kastner, Allamanche, Herre, Cremer, Grossmann, Hellmuth (2002).

²³⁷ See: Jayant, Noll (1984).

Bytes/s und 800 Bytes/s is covered, the default setting corresponding to a rate of approximately 32 Bytes/s. Furthermore, “richer” signatures can be transcoded into “lighter” and more compact signature formats, thus enabling a meaningful comparison between signatures that have been parameterised differently. This *scalability* property permits a high degree of flexibility so that very different requirements can also be satisfied by the same generic data format²³⁸.

IV Applications of Fingerprinting Technology

In the current and future world of multimedia, automatic recognition of content by means of fingerprinting technology has a plethora of attractive applications, some of which are briefly illustrated here:

- *Identification of content and finding associated metadata*: Naturally, fingerprinting technology allows to easily handle unknown audiovisual content (which is *not* annotated by descriptive information) by determining its identity and finding associated metadata. This is an extremely interesting property when trying to benefit from metadata-based services for today’s non-annotated legacy content (such as Compact Discs, VHS video tapes) and works regardless of the kind of media on which the content resides. Note that metadata might also include information on the usage policy associated with a certain piece of audiovisual content.
- *Music sales*: Using this technology, consumers can identify — and possibly order on the spot — interesting content they observed in whatever situation by pressing the identification button on their electronic device. Identification may be performed on PDAs, PCs or via mobile phone.
- *Protection of content-based intellectual property*: Fingerprinting may be employed to find out if and where illegitimate/pirated content is located on the Internet. This is achieved by combining a fingerprinting-based recognition engine with a “web crawler” process which examines the Internet for content and feeds the results through the recognition engine. As a result, a list of “what was found where” can be automatically compiled and used to take-down illicit content.
- *Broadcast monitoring*: Similarly, fingerprinting based systems may be used to implement automated “24 hours per day, 7 days per week” monitoring and analysis of transmitted broadcast programme material. The results can be used e.g. for purposes of media research or simply to verify the accurate transmission of customer’s advertisement spots. Furthermore, analysis of the recovered programme data (“how often was a song/video was played?”) may be utilised to ensure proper compensation of the rights holders for the transmitted content. This type of use has been among the first and very early applications of fingerprinting.

²³⁸ See: Kastner, Allamanche, Herre, Cremer, Grossmann, Hellmuth (2002).

V Digital Watermarking versus Fingerprinting

As can be seen from the previous discussion of the principles of fingerprinting, this technology uses an approach which is clearly different from the one employed by digital watermarking. The following section gives a brief synopsis of the essential characteristics of both types of technologies when used for the purpose of automatic identification of content.

A first obvious reason for the different characteristics of both approaches is rooted in the fact that the process of watermarking implies embedding an information-bearing signal into the content while this is not the case for fingerprinting. This has a number of consequences with respect to the applicability of both technologies to certain usage scenarios:

- In order to employ watermarks it is essential to have access to the content *prior to* its distribution in order to perform the embedding operation. This is not required for the use of fingerprinting-based technology. Consequently, the latter type of technology is also applicable to “legacy content” which has been published before in traditional formats (e.g. Compact Discs or VHS tapes).
- On the other hand, watermarking enables the individual marking of multiple copies of the same works, such that, e.g., it becomes feasible to recover the information on which customer a particular copy was sold to. Fingerprinting-based systems do not provide the capability for such a distinction.
- While watermarking technology always carries a certain risk of introducing perceptual degradations into the content, no such risk exists for a fingerprinting-based approach.
- If it is found desirable at some point to upgrade to a new watermarking scheme with better performance parameters (e.g. higher data rate, robustness or perceptual quality), it is necessary to re-process the entire content database with the upgraded technology and re-distribute the result. An upgrade to an improved fingerprinting system, in contrast, does not require such an effort and may thus be much easier to accommodate.
- In return, watermarking does not require any change of the “receiver” (detector) side if new content items should be included into a service. In the case of fingerprinting, the recognition engine has to be trained to enable detection of the additional content.
- While watermarking does not exhibit a dependency of the computational effort for content identification on the number of different content items to be recognised, the effort for fingerprinting generally increases with the size of the signature database due to the increased search space. (No reference database is needed in the case of watermarking.)

In light of these observations, both approaches show complementary characteristics which can supplement each other very well, depending on the desired application scenario.

VI Conclusions

The concept of automated content-based identification of audiovisual material has received widespread interest recently due to the enormous growth in the amount of available material to everyone and the necessity of efficient handling of such material. The underlying idea for this technology is to perform a similarity search between the unknown (query) item and items stored in a reference database by comparing their condensed representations (fingerprints). The resulting approach for content identification shows properties that are different from — and mostly complementary to — the characteristics of digital watermarking. Both technologies enable a considerable number of very attractive applications in the area of digital rights management and beyond.