DIGIMARC

White Paper

Identifying and Managing Digital Media:
A Technology Comparison of Digital Watermarking and Fingerprinting

9405 SW Gemini Drive, Beaverton, OR 97008 • p +1 800.344.4627 • www.digimarc.com

Identifying and Managing Digital Media

A Technology Comparison of Digital Watermarking and Fingerprinting

Overview

The owners of digital content face a big challenge in the new digital world: how to identify and manage their content in the new and still evolving media life-cycle, in which content now regularly travels unpredictably outside the traditional distribution channels. Content owners need a way to identify and track their content in the new environment as they develop new strategies for monetization. The ability to persistently identify media content at various points in distribution and consumption cycles provides media owners with opportunities for better accounting of their content, optimizing business models, broadening audiences, and enhancing consumer experiences.

This paper defines, discusses, and critiques digital watermarking and fingerprinting, two technologies being used to provide such identification. We will investigate essential characteristics, examine differences, outline example systems, and compare watermarking and fingerprinting on a number of practical criteria.

Introduction

The traditional models of media consumption are being transformed. Photographs, music, movies, and TV shows no longer proceed only along a linear path from content owner to distributor to consumer through pre-authorized, established distribution channels. In the increasingly interconnected digital universe, media content of all kinds disperses at dizzying speeds from sources to consumers along multiple paths through a dazzling variety of digital devices — smartphones, computers, MP3 players, DVD recorders, camcorders, internet servers, etc. — that enable capturing, sharing, time-shifting, clipping, mashing up, and other manipulations of the original content.

In this new culture, consumers increasingly expect access to media whenever, wherever, and however they choose. And consumers have difficulty distinguishing freely available content from protected content that should be paid for. The ease of capturing and sharing digital media, combined with the technical and legal difficulties of controlling and tracking its distribution, often lead consumers to believe, legitimately or not, that they're entitled to enjoy digital content for free. Why pay for something that's readily available on YouTube?

So the traditional models of media monetization are also being transformed, and it's not only unclear how to capitalize on the evolving new media models, it's also unclear how to observe, measure, and characterize the new models so that effective marketing strategies can be developed. Central to the new media models, and a primary cause of difficulty in characterizing and capitalizing on it, is the fact that professionally published content often

loses its *identity* — its connection to its owner — as it moves through non-linear channels and appears in unauthorized venues.

Consider these situations:

- A political parody segment of *Saturday Night Live* was seen on YouTube by a significantly larger audience than viewed the original broadcast.
- Copies of the popular movie *Ratatouille* were available online before the movie was released to theaters.
- There were approximately 350,000 downloads of pre-release tracks of the much anticipated Guns N' Roses album *Chinese Democracy* from a music writer's blog.

Without an effective way to identify content, content owners do not know where their content goes, how it gets there, who its audience is, what devices the audience uses, or what consumers do with the content. With so little known about the journey, destination, and fate of digital media, its owners cannot determine how to establish their presence, engage with consumers, and capitalize on their content. Consumers often come across content by chance or through referrals by a friend or colleague, but without useful contextual information or the value of additional recommendations of similar or complementary content that they may be interested in.

The Opportunity

Lacking the ability to identify content and thereby gain insight into how to develop new media models, content owners lose valuable opportunities to engage consumers — opportunities that are potentially more valuable than ever before. New technologies are enabling new distribution and consumption and new ways to monetize media, but capitalizing on the new models requires that content owners understand the new life-cycle of their content — how, when, where, and by whom it is being consumed.

Benefits of persistent identification to content owners generally fall into three areas, which lead to consumer benefits:

- Better metrics
- This means knowing more about the journey that content takes from content owner to consumer. One benefit of better metrics is ensuring that appropriate credit is given to all who contribute to the value of media — content owners, distributors, aggregators, etc. A second benefit is acquiring detailed and accurate information about media consumers, including precise demographics, preferred delivery channels, advertising effectiveness, and so on.
- Optimal marketing
- Better metrics lead to better marketing. Understanding who consumes media as well as how, when, and where they consume it supports more engaging marketing, targeted effectively at consumers to address their specific interests and increase their response rates.
- Enhanced consumer experience
- Finally, the media consumer also benefits because increased understanding of consumer habits and preferences will enable content owners to provide

experiences that are more entertaining, informative, and useful to the consumer. Content can be tailored to the consumer's preferred devices and locales. More engaging content invites the consumer to receive additional related information and then perhaps to initiate an interaction that benefits both content owner and consumer.

The full impact of these changes is yet to be fully understood, but consider the same three situations described above, only now with content that is identified and traced through its distribution paths:

- The segment of *Saturday Night Live* could be paired with the ads that were broadcast with the original show or with ads customized to the YouTube audience.
- The poor quality, pirated version of *Ratatouille* could be replaced by a trailer for the movie accompanied by links to ticket outlets.
- Full tracks from *Chinese Democracy* could be replaced by samples paired with links to sites where tracks could be purchased and downloaded.

These revised scenarios depend on the ability to identify content moving outside the established channels, an ability that can lead to understanding the new media universe, developing targeted and effective monetization strategies, and enhancing the media experience for consumers. And, as will be illustrated below, if specific instances of content — such as the particular copy of Ratatouille that was leaked — can be identified, this would provide even more benefit to content owners and consumers.

Digital Watermarking and Fingerprinting Technologies

There are two primary technologies currently used to identify content in the new unstructured distribution: ***digital watermarking*** and ***digital fingerprinting***. While both enable content identification, they differ in some significant ways that bear on their appropriateness for different applications.

A ***digital watermark*** is a digital code that can be embedded in all forms of content, imperceptible to people but detectable by computers, networks, and other electronic devices. Conceptually it is analogous to the traditional notion of a watermark on paper, in which a barely perceptible mark is applied during manufacture that establishes the provenance of the paper on later inspection. Similarly, digital watermarks applied to digital content are persistent, staying with the content through manipulation, copying, format conversions, and so on. Digital watermarks are easily detected after distribution, enabling all forms of media and many objects to be given a unique digital identity.

A ***digital fingerprint*** is a unique pattern that identifies content. A fingerprint is derived or computed from selected intrinsic properties of the content. For example, the fingerprints of audio and video content can be derived from salient features extracted from frequencies, timing, color, texture, shape, and luminosity. As with a human fingerprint, the fingerprint of unidentified content must be compared to a database of known fingerprints to identify the original content.

Digital fingerprinting is a form of *pattern* or *image recognition*, and some commercial systems use those terms to describe similar approaches. For this document, any system that identifies content based on its intrinsic properties is considered a fingerprinting system.

Essential Characteristics

Watermarking and fingerprinting have certain intrinsic characteristics that bear strongly on their capabilities and suitability for different applications.

Watermarks:

- Consist of imperceptible data embedded in digital content

  A digital watermark is data embedded in content in a way that is imperceptible to human senses but easily detected and read by computers, networks, and other digital devices equipped with the appropriate software.

- Are applied to content at one or more points before it reaches the consumer

  Watermarks can be applied to content at any point between its creation and final distribution to the consumer. For example, a song can be watermarked when it is burned onto a CD, a TV show can be watermarked when broadcast by a local station, and a movie can be watermarked when it is shown in a theater. It is also possible to embed multiple watermarks in content to identify different points in its path from owner to consumer. For example, the TV show watermarked at the local station can be watermarked again by a set-top box as it viewed by the consumer.

- Carry extrinsic data

  The data encoded into a watermark is called the *payload*. The payload contains *extrinsic* data — data not derived from the content but determined entirely by application requirements. Once extracted, the payload can trigger an immediate action, connect to related information in a database, or both. For example, detection of the watermark in a movie could immediately display the MPAA rating and then connect through a database to the movie's web site.

- Can identify different instances of the same content

  Because watermarks contain extrinsic data and can be applied at different points in the distribution path to the consumer, they permit distinguishing between different instances of the same original content. For example, watermarks applied to TV shows by local stations enable identifying the specific broadcast station of a TV clip uploaded to YouTube. Similarly, watermarks enable differentiation between a movie trailer and the associated full-length movie.

- Do not require a reference database to identify content

  Though many watermark applications access a database, digital watermarks can be useful to identify content without connecting to a database. Wherever content is being monitored, just the presence of a watermark enables immediate feedback to the consumer that the content is owned. And, as mentioned previously, in some cases the payload contains immediately actionable information, such as an MPAA rating. The remainder of the payload is typically a digital identity used to access a database and retrieve additional data related to the content. Since the database can reside either locally or remotely on a server, a network access is not required.

Fingerprints:

- Rely on the uniqueness of content

  Because a fingerprint is a mathematical encapsulation of selected intrinsic properties of content, fingerprints differ only when content differs in the selected properties. For example, if a song's fingerprint is based on its tempo, spectrum, and bandwidth, then different versions of the song are distinguishable only to the extent that those characteristics generate different fingerprints.

- Are the same for all instances of the same content

  A corollary to the above characteristic is that the same content always generates the same fingerprint regardless of where it is consumed. Fingerprints cannot distinguish, for example, the different theaters in which a movie is shown.

- Can be derived from content after it is distributed

  The fingerprint of content can be determined at any time during the life of the content, including after it is distributed to consumers. So, for example, the fingerprint of a movie can be derived after the movie is released to theaters.

- Require a reference database for identification

  Similar to human fingerprints, the fingerprints of all content to be identified must be calculated and entered into a reference database that must be accessible for searching and matching wherever content is being monitored. This reference database is different from the database that contains the additional information related to the content. Depending on the application, the reference database could be local or remote, large or small, relatively fixed or constantly growing, but it must in all cases be accessible.

- Carry no additional data

  Being solely a calculated distillation of the properties of content, a fingerprint contains no additional information. A fingerprint provides identification information after resolution through the reference database.

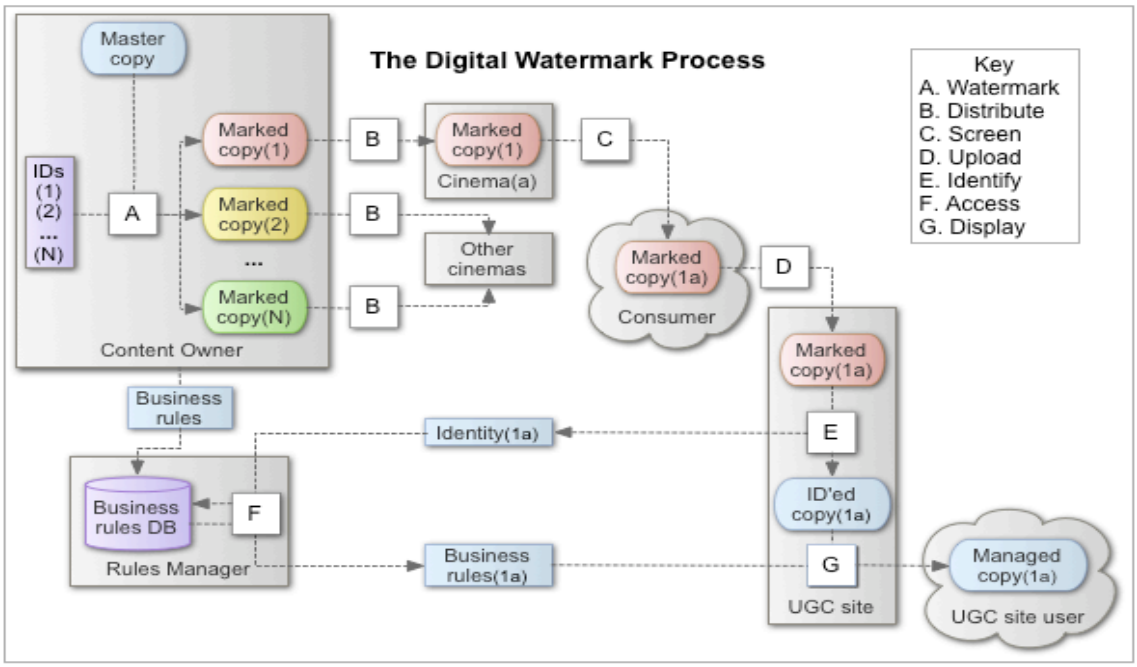
Digital Watermarking and Fingerprinting Systems

For purposes of evaluating watermarking and fingerprinting, it is helpful to understand how systems employing these technologies generally work. It is not feasible to present a single definitive system that covers all situations for either technology, because both watermarking and fingerprinting systems are deployed in a variety of contexts and for a variety of purposes.

This discussion will focus on a representative application — identifying movies uploaded to the web — to illustrate watermarking and fingerprinting systems.

Digital Watermark Process

The following diagram illustrates the steps in watermarking a movie and later identifying the movie when it is uploaded to a UGC (User-Generated Content) web site. The individual steps are explained below the diagram.



A. ***Watermark***: The master copy of the movie is watermarked while still in the content owner's control.

   The watermarks can be embedded by the content owner or a third party. A unique watermark is embedded into each copy of the movie intended for distribution. Each watermark carries a unique identity (represented in the diagram by (1), (2), … (N)), so that downstream it will be possible to determine which distributed copy was found.

B. ***Distribute***: The watermarked copies of the movie are distributed to various cinemas for broadcast.

   The specific means of distribution — whether satellite, the internet, a hard drive, or another digital channel — is not important to this example.

C. ***Screen***: The watermarked movie is screened.

   The movie may be watermarked again during screening to identify the specific theater (represented above by (a)), resulting in multiple watermarks on the same copy (represented by (1a)). It is in fact a requirement of the Digital Cinema Initiative that all new digital cinemas watermark movies during screening to identify the specific venue.

   During screening, someone in the audience surreptitiously records the movie with a small camcorder, capturing a relatively faithful copy of the now doubly-watermarked movie.

D. ***Upload***: By way of their home computer, the consumer uploads the secretly made copy or some portion of it to a UGC Site.

E. ***Identify***: The UGC site reads the watermarks in the uploaded movie and extracts the payloads.

   At this point, it is known from the presence of watermarks that the uploaded movie is copyrighted and the user who uploaded the movie can be notified of this. The payloads contain information that identifies the movie, the distributed copy, the screening theater, and perhaps some immediately actionable information such as the MPAA rating.

   The data extracted from the payloads is communicated to a system called, for this example, the Rules Manager. In most cases the Rules Manager is a remote system, not resident on the UGC site, although this is an implementation issue.

   The movie's identity is the primary item extracted from the payload, but the data identifying the screening theatre enables better understanding of how and where movies are being duplicated and ultimately contributes to deterring piracy.

F. ***Access***: The Rules Manager accesses business rules and other information associated with the movie.

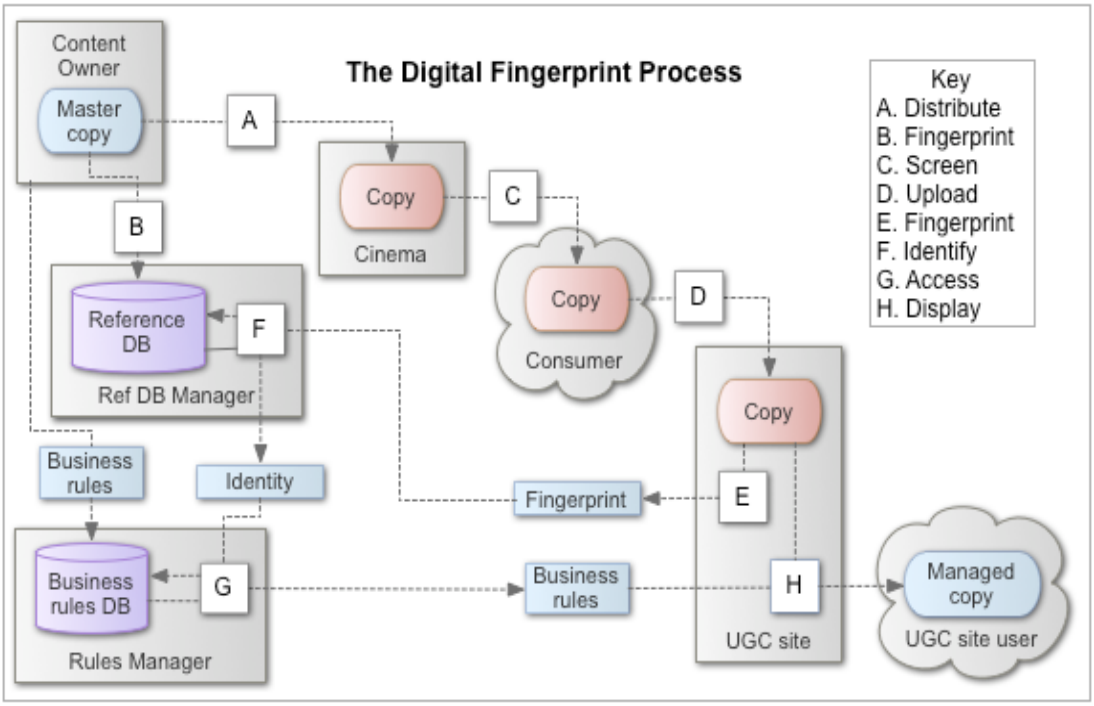
   The Rules Manager may or may not be the Content Owner, but in any case the business rules originate from the Content Owner. The rules are returned to the UGC site and the movie is now reconnected to the Content Owner.

G. ***Display***: When any UGC site user chooses to view the identified movie, the UGC site applies the business rules to determine how to display the movie.

   The rules permit the Content Owner to specify a broad range of controls on viewing the movie: Display could be prohibited entirely; a trailer could be shown; if the location of the user is known, the UGC site could present the times and locations of local screenings of the movie; and so on.

Digital Fingerprint Process

The following diagram illustrates the steps in applying fingerprinting to identify a movie uploaded to a UGC web site. The individual steps are explained below the diagram.



A. ***Distribute***: Copies of the movie are distributed to cinemas for broadcast.

B. ***Fingerprint***: The movie's fingerprint is calculated and entered into the reference database.

   The movie's fingerprint can be determined at any time before it is needed for identification (step F). And the fingerprint need not be calculated from the master copy of the movie — any sufficiently high-quality copy will do.

C. ***Screen***: The movie is screened.

   During screening, someone in the audience surreptitiously records the movie with a small camcorder, capturing a relatively faithful copy of the movie.

D. ***Upload***: By way of their home computer, the consumer uploads the secretly made copy or some portion of it to a UGC Site.

E. ***Fingerprint***: The fingerprint of the uploaded copy is calculated.

   Calculating the fingerprint of the uploaded copy reveals nothing about the movie's identity, copyright status, rating, etc.

   The fingerprint is communicated to a system for comparison to the reference database. In most cases the reference database is on a remote system, not resident on the UGC site, although this is an implementation issue.

F. ***Identify***: The calculated fingerprint is matched against the reference database to determine the movie's identity.

   As will be explained below, fingerprinting is an inherently statistical process the accuracy of which depends on a variety of factors. This means it is not necessarily guaranteed that the correct match will always be found for a fingerprint.

G. ***Access***: The Rules Manager accesses business rules and other information associated with the movie.

   This step is similar to the ***Access*** step (F) in **The Digital Watermark Process** above, except that a fingerprinting system cannot distinguish between different instances of the same content. So the Rules Manager applies the same business rules to the content regardless of its distribution channel.

H. ***Display***: When any UGC site user chooses to view the identified movie, the UGC site applies the business rules to determine how to display the movie.

   This step is similar to the ***Display*** step (G) above, with the caveat that the business rules cannot be applied to individual instances of content in a fingerprinting system.

There are, of course, many variations on these processes and implementation considerations have been largely ignored in this explanation. But for purposes of further discussing and comparing the two technologies, this overview illustrates the significant points.

Digital Watermarking and Fingerprinting Comparison

This section compares digital watermarking and fingerprinting on a number of criteria to provide a better understanding of the advantages and disadvantages of each.

Accuracy

Accuracy is defined here as the certainty that the identification of content is correct. Correct identification includes the avoidance of both *false positives* — incorrectly identifying content — and *false negatives* — incorrectly failing to identify content. Of these two, false positives are more damaging, because wrongly claiming ownership of content would, to everyone's detriment, rapidly undermine consumer confidence in the system.

Watermarking

Watermarking identifies content based on extrinsic or independent data embedded imperceptibly in the content. Because watermarking is based on established mathematical models of digital communication, accuracy in watermarking systems is a design parameter that is independent of the size or composition of a reference database. Because identification by watermarking doesn't rely on such a database, it is not subject to the potential for database matching errors. The system's accuracy depends only on the nature of the watermarking algorithm, the content itself, and the degree of content degradation.

In practice, this means that accuracy in watermarking system is a parameter that can be specified with little uncertainty, regardless of the volume of content.

Digital watermarking systems have been deployed in real-world applications for over 10 years and have empirically demonstrated a high level of accuracy across billions of units of watermarked content. The definitive nature of watermark identification is a particular advantage when the volume of content, number of instances, or usage models (e.g. degree of consumer manipulation) cannot be determined beforehand. In such situations, accuracy remains constant, a very beneficial result for content owners.

Fingerprinting

Fingerprinting identifies content based on its intrinsic properties, so a fingerprint can always be calculated for an instance of content, even if the instance is degraded. The result of matching an unknown fingerprint to the reference database is dependent on several variables, such as the quality of content from which reference fingerprints are derived, the size of reference database, and the number of perceptually similar references already contained in the database.

The accuracy of a fingerprinting system is not a specifiable parameter as it is in a watermarking system. A fingerprinting system's accuracy is usually determined empirically, by observations in practice. While fingerprinting systems have demonstrated accuracy up to 98% in certain existing applications, the scale of existing applications is relatively small, so the accuracy of fingerprinting on a large scale is yet unknown. Accuracy is dependent on the variables mentioned previously – database size, etc. – and how these variables are constrained and how they interact has not been established for large scale fingerprinting systems.

One implication of this is that the accuracy of fingerprinting systems is difficult to determine over time and increases in volume. This issue has sometimes been managed by techniques such as offering users the option of choosing the best of a set of closely matching candidates or narrowing the comparison domain through keyword filtering.

Efficiency

***Efficiency* is the speed with which a determination is made on unidentified content. Efficiency has three components in watermarking and fingerprinting systems:**

- Extracting the watermark or calculating the fingerprint
- Transmitting the watermark or fingerprint to a remote database for identification
- Processing the watermark or fingerprint to identify the content

In practice, the time to extract a watermark is comparable to the time required to calculate a fingerprint, so the first component can be disregarded for purposes of evaluation. Because the transmission times for fingerprints are only slightly greater than for watermarks, transmission times can also be ignored. So the comparison of efficiency for watermarking and fingerprinting systems is primarily based on the time required to make a determination of identity from the watermark or fingerprint.

Watermarking

In a watermarking system, content is generally identified by using a portion of the payload as an index into a database. Indexing into a database is a single operation the cost of which

is easily and accurately determined and essentially constant regardless of the size of the database. So the efficiency of a watermarking system is known and constant regardless of the volume of content.

Fingerprinting

In a fingerprinting system, content is identified by searching the reference database for the closest match to the fingerprint of the unknown content, so the search time determines the efficiency of the system. Database searching is an operation the cost of which typically increases with the size of the database. Since most of the existing fingerprinting systems are relatively small in scale, database search time has not been a major issue, but as the volume of content grows and reference databases increase in size, search times could increase measurably.

With an increase in size; however, search times will be dependent on the resources applied to the search operation. For example, additional hardware resources, improved search algorithms, and faster processors can all be applied to control search times. But it is difficult to evaluate at a system's inception how increasing volumes of content can be balanced by additional database and computing resources.

Complexity

Complexity is defined here as the level of difficulty in developing and maintaining a reliable system.

Watermarking

Watermarking systems are conceptually straightforward, essentially consisting of three steps:

- Embed watermarks in content.
- Extract watermarks from content.
- Access content-related information.

The mathematics of embedding and extracting watermarks is fairly complex, but it has been thoroughly studied and widely used for over ten years. For most applications, the algorithms are well developed and well tested.

It is a simplifying factor in watermarking systems that identifying content is an algorithmic process rather than a database search. Given an adequate sample of content and a desired level of accuracy, either a watermark can be read or it can't. If it can be read, identification of the content is very accurate (see **Accuracy** above); if the watermark can't be read, no identification is possible and none is attempted. Watermarks are typically embedded repeatedly throughout content, so any intact sample usually yields a watermark.

Accessing content-related information typically requires using the watermark's payload to index into a database to determine how to handle the identified content. And, as discussed previously, indexing into a database is a simple operation, regardless of the size of the database.

Fingerprinting

Fingerprinting systems are similar in data flow to watermarking systems but contain an additional step. The basic steps in a fingerprinting system are:

- Calculate the fingerprints of known content to build a reference database.
- Calculate the fingerprint of unknown content.
- Match the unknown fingerprint against the reference database.
- Access content-related information.

The third step, matching an unknown fingerprint, is the additional step with the greatest potential to add significant complexity. While searching databases is a well-understood operation, two factors contribute to its being more complex in a fingerprinting system: first, growth in content volume; second, the inexact nature of fingerprint matching.

Consider growth first. As the volume of fingerprinted content increases, the database storage requirements and the processing power for searching the database both increase. The absolute number of search requests also increases with volume, leading to a larger overall load on the reference database.

Various techniques have been used to manage a growing reference database. One approach is to apply multiple processes to perform simultaneous, parallel searches of the reference database. Another, similar approach is to distribute multiple copies of the reference database to the points where content is to be identified, thereby distributing the search load.

The upshot of this is that a match occurs when the unknown fingerprint is sufficiently close to a reference fingerprint. *Sufficiently close* is a system parameter that may require tuning as the system increases in size and the database becomes denser with more reference fingerprints. Modifying the closeness parameter can affect both efficiency and accuracy in ways difficult to predict.

The first step in the fingerprinting process — the calculation of fingerprints and construction of the reference database — can also increase the complexity of a fingerprinting system in some cases.

Cost

The *cost* of a system includes the up-front costs of implementation plus the long-term costs of use, maintenance, and expansion. To be considered are not just objective costs like computer resources but also more subjective factors such as system complexity, which impose difficult-to-measure costs in additional development and maintenance.

For purposes of comparison, the costs of developing and distributing software readers to extract watermarks or to calculate fingerprints are comparable, so those costs are ignored here.

Watermarking

Watermarking is a well-established technology that has been applied to very large systems, so its costs are well understood and predictable. A watermarking system entails an up-

front investment in that it requires the ability to embed watermarks to be integrated somewhere in the system of content production and distribution. This up-front investment is a fixed cost that is amortized over the volume of content watermarked. If the deployed embedding capacity eventually becomes saturated, an additional incremental investment in embedding technology is required, and this fixed cost will likewise be amortized over the next mass of content. So the cost of embedding watermarks is a step function that increases in predictable increments as volume increases.

The downstream cost of watermarking consists primarily of updating the identity database with new content identities and associated information. As the volume of unique instances watermarked increases, minimal cost is incurred in updating the identity database.

From a development and maintenance perspective, watermarking systems are relatively simple, both in overall concept and in the complexity of individual steps. For example, as mentioned above, the database operation to identify content is a simple indexing operation. And the complexity of a watermarking system is constant as the system grows in size.

Fingerprinting

Costs of fingerprinting systems can be difficult to establish with certainty, particularly as systems grow in scale. With respect to up-front investments, fingerprinting systems have a lower barrier to entry, because fingerprinting does not require integration into the content distribution process. There is still some up-front costs in calculating fingerprints to build the reference database.

The predominant cost of a fingerprinting system is incurred in maintaining and searching the reference database. The discussion so far has revealed that, as the system's volume increases, the reference database grows in size and becomes more difficult to manage.

It is difficult to make more precise statements about the costs of large-scale fingerprinting systems because of the lack of experience with large systems. But increases in volume are typically accompanied by growth in a system's size, complexity, and cost.

Scalability

Scalability is how well a system performs as its size increases. For content identification and management systems, performance consists of accuracy, efficiency, complexity, and cost — the topics of the previous several sections — so scalability represents these factors taken together.

Watermarking

From the discussions above, it follows that watermarking systems are highly scalable:

- Because the identity of any instance of content is determined solely by the ability to read that content's watermark, the overall volume of content has no bearing on the accuracy of the system.
- The efficiency of a watermarking system is determined by three factors — embedding the watermark, extracting the watermark, and indexing into a database — that are all constant regardless of the volume of the system.

- The complexity of watermarking systems is relatively low and does not grow with system volume.
- The absolute cost of a watermarking system increases with volume in a predictable way. The relative cost per instance of content identified actually decreases because the up-front costs of the system are amortized over the volume of content.

Fingerprinting

Based on the same factors, it is reasonable to conclude that fingerprinting systems may not scale as well:

- The accuracy of fingerprinting in large scale systems is not yet well established. It depends on the interactions of several variables that can be affected by content volume with unpredictable results.
- A fingerprinting system's efficiency is primarily dependent on the speed of searching the reference database. The relationship between increased resources and efficiency is not well understood for large systems.
- Fingerprinting systems increase in complexity as the system grows in ways that are difficult to anticipate and have not been investigated in any depth.
- With increasing system volume, the cost of a fingerprinting system may increase in an unforeseeable manner.

Legacy Content

Legacy content is content that has been distributed and is in consumers' possession before a system is in place to identify and manage it.

Watermarking

Watermarks must be embedded in content at some point during the production and distribution process, before the content is in the hands of consumers. Once consumers possess content, it is traditionally beyond the reach of the watermarking system. This means that, in general, watermarks cannot be used to identify legacy content. However, if content re-enters the content owner's arena, it can then be watermarked. For example, if an unmarked movie trailer is posted to a web site, it can then be watermarked for future identification and tracking. Also, if content originally released in a non-digital medium such as film or vinyl (for records) is re-mastered and re-released in digital form, it can be watermarked during the re-release process.

Fingerprinting

Unlike watermarks, fingerprints are not external information embedded in content but are representations of the content itself. As a result, a fingerprint can be calculated from any instance of content, including an instance that is already in possession of consumers. For example, fingerprints can be calculated for movies originally released on film and then used to identify versions of those movies that have been captured by camcorder and uploaded to web sites.

Specificity

Specificity in the context of a content identification system is its capability to identify specific instances among multiple copies of the same content. For example, if a movie trailer is posted to a web site, is it possible to identify which distributed copy of the trailer it is or even in which movie theater the trailer was screened?

Watermarking

A unique watermark can be embedded in any copy of content at any stage in its distribution, including, as in the example above, at the screening of a movie trailer in a particular theater. Watermarks provide the capability to definitely identify instances of content to any desired level of specificity. Watermarks permit distinguishing the different copies and even different screenings of the movie trailer, versions of the same ad appearing in different magazines, the soundtrack versus the CD version of a song, etc.

Revisiting the examples from the Introduction highlights the benefits of specific identification of content:

- Knowing the broadcast source of the *Saturday Night Live* clip helps understand the show's demographics.
- The ability to identify the source of the leak of *Ratatouille* helps deter piracy.
- Identifying the music writer who prematurely published Guns N' Roses tracks could help build a legal case against the writer or, alternatively, lead to developing a cooperative business model with the writer.

Fingerprinting

Being calculated from the properties of content, fingerprints of two different instances of the same content are identical. They cannot then be used to distinguish one instance of content from another instance of the same content. The movie trailer posted to the web is indistinguishable from all other copies of the same trailer and a short scene from a trailer cannot be distinguished from the same scene appearing in the movie itself. Similarly, copies of an ad appearing in different magazines and different instances of a music file are indistinguishable by fingerprinting.

Summary

The following table summarizes the above comparisons of digital watermarking and fingerprinting.

Summary of Digital Watermarking and Fingerprinting

	Watermarking	Fingerprinting
Definition	Digital information embedded in content, imperceptible to people but detectable by digital devices	Represents content by a mathematical calculation based on intrinsic properties of content
Identification mechanism	Payload carries identity	Fingerprint matched against a reference database
Accuracy	0. A specifiable design parameter 1. Historically demonstrated in large systems 2. Rare false positives 3. Constant over system growth	0. Accurate in current, small-scale systems 1. Potential inexact comparisons may use approximations
Efficiency	0. Defined by database indexing operations 1. Constant over system growth	0. Defined by database searching operations 1. Potentially improved by increased investment in resources
Complexity	0. Well understood algorithms 1. Simple database operations 2. Algorithmic identification	0. Possible database optimization 1. Potential inexact database matches 2. Database update issues
Cost	0. Up-front cost for embedding infrastructure 1. Increases overall in predictable ways with system growth 2. Decreases per instance of content with system growth	0. Lower barriers to entry for content owners 1. Low up-front cost for reference database 2. Increases unpredictably with system growth
Scalability	0. Accuracy constant 1. Efficiency constant 2. Complexity constant 3. Cost decreases per instance of content	0. Efficiency and cost are currently less predictable
Legacy content	0. Not applicable to legacy content unless re-released	0. Applicable to legacy content
Specificity	0. Identifies individual instances of same content	0. Cannot identify individual instances of the same content

Conclusion

In summary, content owners in the digital age clearly face the challenge of identifying and managing their content as it quickly travels outside traditional distribution channels. A way is needed to identify and track content in the new environment to support new business models and strategies for monetization. The ability to reliably identify digital content during distribution and consumption provides content owners and distributors with opportunities to achieve better accounting of their content, optimal marketing and monetization, and enhanced consumer experiences.

This discussion has presented two technologies, digital watermarking and fingerprinting, with the capability of addressing this challenge. Both technologies can identify media content, but they operate differently to do so.

The primary advantages of watermarking are:

- The association of extrinsic data with content to provide a unique and persistent digital ID
- The ability to identify individual instances of content
- Significant history in real applications
- Well understood, consistent system behaviors with respect to accuracy, efficiency, complexity, and cost
- Scalability

And the primary advantages of fingerprinting are:

- Applicability to legacy content
- No modification of distributed content
- Low barriers to adoption of a system for small scale implementations

While this paper has compared and contrasted the two technologies as used in independent systems, they are not mutually exclusive. Digital watermarking and fingerprinting can be complementary and coexist smoothly in a single implementation. In some cases such a hybrid system may likely provide the best overall solution, for example to identify TV clips uploaded to UGC sites. Such a hybrid system may correctly identify more content than either technology would if applied alone. In the real world of consumers who sample, copy, convert, and otherwise mangle digital media, watermarking and fingerprinting can complement each other to achieve the desired results.

About Digimarc

Digimarc Corporation (NASDAQ:DMRC), based in Beaverton, Oregon, is a leading innovator and technology provider, enabling businesses and governments worldwide to enrich everyday living by giving persistent digital identities to all forms of media and many other objects. The company's technology enables solutions for enhancing traffic safety and national security; deterring fraud, counterfeiting and piracy; and facilitating new digital media distribution and monetization models that provide consumers with more choice and access to content when, where and how they want it. Digimarc has an extensive intellectual property portfolio in digital watermarking, media identification and management, and related technologies. Digimarc develops solutions, licenses its intellectual property, and provides development services to business partners across a range of industries.

We're ready to help you

For more information, visit www.digimarc.

Digimarc Corporation
9405 SW Gemini Drive
Beaverton, OR 97008
T +1.800.344.4627 • F +1.503.469.4777
www.digimarc.com

Copyright © 2010 Digimarc Corporation. Digimarc and the Digimarc logo are registered trademarks of Digimarc Corporation. All rights reserved. All other trademarks are the exclusive property of their respective companies.