

Watermarking and Ownership Problem: A Revisit

Husrev T. Sencar
Polytechnic University
Brooklyn, New York
taha@isis.poly.edu

Nasir Memon
Polytechnic University
Brooklyn, New York
memon@poly.edu

ABSTRACT

Watermarking technologies have been envisioned as a potential means for establishing ownership on digital media objects. However, achievable robustness and false-positive rates of the state-of-the-art watermarking techniques raise doubts about applicability of watermarking to ownership problem. With this perspective, we address the security weaknesses common to most watermarking techniques and assess the role of watermarking in construction of ownership assertion systems. We identify the requirements of a watermarking based ownership assertion system. Also, we provide a basic functional outline of a practical version of such a system and identify its potential vulnerabilities. To mitigate these vulnerabilities, we aim at reducing the false positive rate of the watermark detection scheme. For this purpose, we propose embedding multiple watermarks as opposed to single watermark embedding while constraining the embedding distortion. The crux of the proposed method lies in watermark generation which deploys a family of one-way functions. We incorporate the multiple watermark embedding idea with the additive watermarking technique [1] and present results to illustrate the potential of this approach in reducing the false-positive rate of the watermark detection scheme.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Security

Keywords

watermarking, ownership, counterfeit ownership, ownership deadlock, theft of ownership, additive watermarking

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DRM'05, November 7, 2005, Alexandria, Virginia, USA.
Copyright 2005 ACM 1-59593-230-5/05/0011 ...\$5.00.

1. INTRODUCTION

The ownership rights over intellectual property have long been recognized and clearly defined by rule of law. These laws, in common, grant the *creator* (owner) exclusive rights to copy or distribute a protected form of their *property*, thereby protecting its unauthorized use. Essentially, this requires means of controlling how the protected work is utilized by others. Development of technologies that will enable establishing ownership on intellectual property has been the focus of watermarking research. In this context, watermarking techniques are intended to serve the purpose of an invisible tag to identify an object, control and manage its use, and trace its point of dispersal. However, the deficiencies in the design of watermarking methods have limited their applicability to establishing rightful ownership of digital objects (*ownership problem*) in a practical setting.

In most applications of watermarking the main concern has been the robustness against attacks. Yet, another very important and often neglected component of the watermarking system design is the incorporation of security considerations to the design process. Watermarking based approaches to ownership problem have not been convincing mostly due to lack of security awareness. In terms of security, the inadequacy of prevailing design paradigm in tackling the relevant ownership issues were mainly due to an incomplete assessment of the threat model, which requires a thorough evaluation of how proposed schemes can be deployed in unintended and malicious ways to impede the goals and purpose of watermarking. From this standpoint, many of the problems studied in the context of security and cryptography are similar to the ones encountered in the ownership problem, therefore, it is essential for watermarking techniques to merge signal processing methods with the vast body of knowledge in these fields. In this regard, there have been many approaches proposed in the literature applying the existing solutions to various aspects of the ownership problems [2][3][4][5][6][7][8] [9].

The deployment of watermarking techniques for protecting owners' rights requires efficient protocols that accurately and uniquely describe and specify the use of watermarking techniques. Although there are various formalisms proposed for investigating and analyzing protocols to see whether they are prone to design flaws, designing secure protocols that are immune to malicious use remains to be a difficult task. Furthermore, since watermarking techniques depend on statistical methods, erroneous decisions are possible, and as a consequence they are also prone to various attacks engineered to exploit such weaknesses. The attempts to eliminate the

security weaknesses of watermarking techniques have usually taken the path of simply fixing each vulnerability that comes to light. However, the proposed solutions were either unable to identify the full extent of the problem and, therefore, some other forms of deficiencies were inherent or they did not lead to a practical model for an ownership assertion system. As a consequence, the possibility of realizing a watermarking based ownership system has been a major concern.

In this paper, we attempt to provide a unified framework to solve the ownership problem in the light of the previous studies and results. For this purpose, we discuss the requirements of an ownership assertion system and contemplate the role of watermarking techniques in construction of such systems. With this perspective, we outline a sketch of a practical ownership assertion system. In this context, we address several security weaknesses of the watermark techniques and identify the challenges to be met for successful deployment of them. Among the most daunting challenges are the robustness limitations and high false-positive rates. Devising robust watermarking techniques have been the primary focus of the research community from the very beginning as it is central to almost all watermarking applications. Because of these efforts, today, watermarking techniques are able to offer a degree of robustness against a *limited* attacker. For the scope of this paper, we rely on the presence of such watermarking techniques. On the other hand, false-positive rate of a watermarking technique is of utmost importance in the design of ownership assertion systems as it is the basis for various attacks. To render subsequent attacks more difficult, we propose multiple watermark embedding, as opposed to single watermark embedding, under constrained embedding distortion. We analyze the effectiveness of this approach and apply it to additive watermark technique of [1]. The results show that at a fixed embedding distortion level embedding multiple watermarks yield lower false-positive rates as compared to embedding a single watermark.

In the text, we use the following notation. Upper-case letters in *italic* and calligraphic typeface denote vectors and sets, respectively, e.g. $X = \{x_1, x_2, \dots, x_n\}$ and $x_i \in \mathcal{X}$. The operator $|\cdot|$ indicates the cardinality of a set or the length of a vector and the notation \mathcal{X}^n represents a length- n sequence with elements in \mathcal{X} , e.g., \mathbb{R}^n and $\{0, 1\}^n$.

2. BASIC DEFINITIONS

Throughout this paper, we restrict ourselves to the use of watermarking schemes as a means to establish ownership and resolve ownership disputes. With this perspective, we provide a description and generalized formulation of watermarking schemes. In this setting of the problem, an author (owner) creates a digital object (*i.e.*, document, image, audio, video) which needs to be protected from intellectual piracy. To achieve this, the owner purposefully modifies the object by embedding an invisible *signature* (*watermark*) and then makes this version of the object publicly available so that, when needed, the presence of the watermark can be shown as a proof of ownership. Herein, the unpublished version of the newly created object will be referred to as *cover-object* or *original*, and it will be denoted by C . The watermark, and the modified and published version of the *original*, called the *embedded-object* or *watermarked-object*, will be denoted by W and E , respectively. Also, the set of all cover-objects and watermarks will be indicated by \mathcal{C}

and \mathcal{W} , respectively. It should be noted that the set \mathcal{C} also contains all embedded-objects since E and C need to be of the same type.

A watermarking system designed to serve ownership claims and proprietary claims have three major components: *watermark and key generation*, *embedding* and *detection*. The watermark and key generation is composed of two algorithms. The watermark generation algorithm yields a data string W which will be used to associate an object to its owner. On the other hand, key generation is necessary for the functions that are intended to enhance the security aspects of the watermarking system by incorporating cryptographic primitives and introducing asymmetry into watermark insertion and extraction. The output of this algorithm is a pair of keys $(K_{\mathcal{E}}, K_{\mathcal{D}})$ (from a finite set \mathcal{K}) which are not necessarily distinct and when they are not, the output will be the key K .

The second component is concerned with watermark insertion. This is realized by an embedding function (embedder) \mathcal{E} which embeds the watermark W in cover-object C under key $K_{\mathcal{E}}$ yielding the embedded-object E as

$$E = \mathcal{E}(C, W, K_{\mathcal{E}}) \quad (1)$$

where E and C are *perceptually* very similar. The last component of a watermarking system extracts the watermark. Depending on the design of the system a detection function (detector) \mathcal{D} might either extract a watermark or check the presence of a particular watermark W in a given embedded-object E (or in a possibly modified version \hat{E}) under key $K_{\mathcal{D}}$. Furthermore, watermarking systems are classified into two based on the use of cover-object in extraction of the watermark, namely *blind* and *non-blind* watermark detection schemes. When \mathcal{D} is to extract a watermark from the object \hat{E} in a *blind* manner, the output is expressed as $\hat{W} = \mathcal{D}(\hat{E}, K_{\mathcal{D}})$. However, with non-blind detection, where existence of the W in \hat{E} is in question, \mathcal{D} is forced to output a boolean value to indicate the presence or absence of W as

$$\mathcal{D}(\hat{E}, W, K_{\mathcal{D}}) \in \{true, false\}. \quad (2)$$

In the case of non-blind watermark generation, the detection function, (2), take also as input the cover object C . It should also be noted that not all embedder/detector designs require keys in their operation. The above formulation, with the removal of keys, applies to those schemes as well.

3. OWNERSHIP PROBLEM

The most fundamental functionality of an ownership assertion system is to provide the owner of an object with the capability of asserting ownership rights on all objects derived from the original. Essentially, this requires means to generate a protected version of the original object so that when an act of piracy is committed the owner may commence a legal action. For this purpose, we assume the owner embeds a watermark in the cover-object using a robust embedding scheme and releases the resulting embedded-object instead of the original. Accordingly, in case of a dispute the owner may present the unpublished original and show the presence of the embedded watermark as a proof of ownership. In this context, an *ownership problem* emerges when for a given object rightful ownership cannot be resolved.

An ownership problem may be in one of the three forms.

Ownership deadlock: The pirate is able to provide an ownership proof that is as conclusive as the actual owner's proof. Therefore, ownership cannot be established and a deadlock arises.

Counterfeit ownership: The pirate is able to provide an ownership proof that is more convincing than that of the actual owner. Consequently, the pirate may proclaim counterfeit ownership over the object whose ownership is in question.

Theft of Ownership: The pirate obtains an embedded-object and embeds a new watermark in it (pretending it is a cover-object). Hence, the pirate claims ownership on a variant of a cover-object whose actual owner is oblivious to what is occurring.

A solution to ownership problem requires that when two or more parties are involved in a dispute, where all parties claim to hold ownership rights of a particular object, the true owner be identified reliably. This problem is further exacerbated in cases where the true owner is not involved in the ownership dispute and all claimants have to be rejected.

In the field of watermarking the main research focus has been the design of embedding/detection techniques that can survive very sophisticated attacks which subject the embedded-object to a variety of signal processing operations or cryptographic attacks so that the extractor is no longer able to find traces of embedded watermark. However, the ownership problem is mainly due to a more malicious and effective class of attacks, called *protocol attacks*, and even the most robust watermarking techniques may be vulnerable to this type of attacks if they are not designed and used properly. There are two types of protocol attacks, namely *copy attacks* [10] and *ambiguity attacks* [2]. The main idea of a copy attack is to copy a watermark from an embedded-object to any other object and the objective of ambiguity attack is essentially to detect a watermark that was inherently present (not through embedding) in an object. In terms of their impacts the two types of protocol attacks are very similar; however, the vulnerabilities they are exploiting are different. In this regard, success of copy attacks rely on the fact that statistical properties of the watermark can be distinguished from those of the cover-object. Therefore, copy attacks can be circumvented by signal processing measures. On the other hand, applicability of ambiguity attacks depends primarily on the false-positive rate of the embedding/detection scheme, and they cannot be completely avoided unless the false-positive rate is reduced to arbitrarily low values. However, due to the variety of attacks the embedded-object may undergo prior to watermark detection, designing robust embedding/detection techniques with very low rates of false-positive remains to be a challenging task. Therefore, our main concern in this work is the ambiguity attacks.

Craver *et al.* [2] introduced the first realization of an ambiguity attack, called *inversion attack*. The basis of the attack lies in the notion of invertibility of embedding operation. In an inversion attack, the attacker obtains an embedded-object E and finds a watermark W^* , through a brute-force search, that can also be detected in E . Hence both $\mathcal{D}(E, W) = true$ and $\mathcal{D}(E, W^*) = true$ hold. Then, the attacker generates a fake original C^* by de-embedding (subtracting) the watermark W^* from E . As a consequence, the watermark W can be reliably detected in C^* as well as

W^* in C (unless $W = W^*$). Thus, the attacker causes a deadlock by making it possible to link the embedded-object to two distinct originals unequivocally, at the complexity of finding a valid watermark W^* .

3.1 Approaches to Ownership Dispute Resolution and Authorship Proofs

The first step towards a framework that also include authorship proofs and dispute resolution capabilities was taken by Craver *et al.* [2] through imposing the *invertibility* requirement on the embedding/detection scheme. This initiated a serious of work that aim at devising *non-invertible* schemes that are mainly built on conventional (embedding/detection) techniques. In order to achieve *non-invertibility*, Craver *et al.* [11] proposed to include one-way (trapdoor) functions along the path of watermark generation, so that it is not possible to reverse the process. Inspired by [2], Qiao *et al.* [12] proposed the use of standard encryption functions for watermark generation. In their construction, the watermark is created by encrypting some information derived from transform coefficients of the cover-object under a predetermined key selected by the owner, and ownership verification requires both the original and the key. In a similar manner, Zeng *et al.* [13], considering additive embedding, imposed two limitations on the watermarking technique. First limitation requires that C not be used for watermark extraction because extracting the watermark from the difference of C and \hat{E} will not yield a true false-positive rate of ownership claim since detector is unable identify the cover-object. The second limitation provides that the watermark cannot simply be a random signal for which they utilized a one-way function. Another early proposal was to use time-stamps to generate the watermark [14].

In [15] and [16], Ramkumar *et al.* showed that if the false-positive rate of the underlying embedding/detection scheme is high, the cryptographic constructions deployed in watermark generation does not provide a basis for establishing ownership because the pirate does not need to reverse engineer the watermark generation process to obtain a valid watermark to claim counterfeit ownership. In essence, the pirate uses the *diffusion* property of cryptographic constructions to his advantage by generating many watermarks via introducing insignificant changes to a severely altered (but perceptually intact) version of embedded-object E , \hat{E} . When the number of resultant watermarks is in the order of the false-positive rate, it is very likely that one of the watermarks will yield a satisfactory detection statistic. The pirate can now designate the slightly modified version of \hat{E} , that yielded the particular watermark, as his fake original and claim counterfeit ownership. Therefore, non-negligible rates of false-positives poses a challenging problem to resolving rightful ownerships. To render counterfeit attacks of this nature more difficult, Ramkumar *et al.* [16] proposed a means to lower the false-positive rate of the scheme by defining a new detection statistic. For this, they required blind detection of the watermark as in [13] and imposed an added constraint that requires the independence of the watermark with the original. Accordingly, the detection statistic is obtained as a combination (*i.e.*, weighted difference) of two statistics where the first one indicates the presence of the watermark in the embedded-object and the second term signifies the non-existence of the watermark in the cover-object. Hence, to claim ownership on the true original, the

pirate has to ensure that the fake watermark exhibits *strong presence* in the true original and at the same time *no presence* in his fake-original. This is very difficult to achieve. Since the true original and the fake one are still expected to be very *close*, a random sequence (*i.e.*, pirate's watermark) that can be detected in any of the two is very likely to be detected in the other as well. For the case of low false-positive rates, on the contrary, Li *et al.* [17] formally proved that non-invertibility can be achieved and showed that provably secure techniques are present.

In order to get around the limitations of embedding/ detection schemes in resolving ownership disputes Katzenbeisser *et al.* [4] proposed an alternate approach which has been further developed by Adelsbach *et al.* [5]. In their construction, the computation of the watermark is hardened by incorporating digital signatures of a trusted party rather than deploying mechanisms to achieve *non-invertibility*. The watermark is generated by concatenating cryptographic *signatures* corresponding to various *messages* (*i.e.*, original, key, payload, etc.) into a known pattern. In this case, for a pirate to claim ownership (of a cover-object) the extractor should detect and verify validity of pirate's watermark. (That is, the watermark has the right pattern and contain valid signatures.) As compared to schemes described in [11] [16], this approach restricts brute forcing capability of the pirate since computation of each watermark requires querying the trusted party. Assuming the attacker is not allowed to query the trusted party and the signature scheme is secure then this scheme is *provably* secure against counterfeit ownership attacks. However, in a practical setting the trusted party has no way of distinguishing a potential attacker from a legitimate user, therefore, a passive attacker assumption is not realistic. On the other hand, when the attacker is able to make unlimited queries to the trusted party, the problem reduces to one discussed above and the complexity of an attack depends on the false-positive rate of the embedding/detection scheme. A solution suggested in [5] to circumvent unlimited querying problem is to let trusted party keep records of the messages which have been previously signed and to deny response to those messages when they are repeatedly queried. However, this requirement on the trusted party may serve as a possible bottleneck. Because when the trusted party makes a decision by checking if the exact same messages appeared before, it becomes vulnerable to a variant of the attack described in [16]. Therefore, the trusted party cannot reject signing messages on the basis of exact match but should also consider *similarity* between the submitted messages which is not a trivial task.

Adelsbach *et al.* [3] were the first to point out a problem common to all of the above schemes. They recognized that previous approaches to proving rightful ownership achieved only dispute resolution. That is, only if the actual owner is involved in an ownership dispute, those schemes can guarantee that the winner is the owner. Otherwise, the outcome is not conclusive. Simply, there were no mechanisms in place to determine whether the object in question was initially watermarked by the actual owner and later one of the disputants introduced another watermark in the embedded-object relying on the assumption that the actual owner will not be informed of the dispute and be able to prove presence of his watermark therein. Therefore, a true dispute resolving scheme should not resolve the dispute in favor of one of the disputants if the actual owner is not involved in the dispute.

In their model for ownership proofs, owners are required to register their work (in order to claim ownership on it) at a center which in return generates a variant of the work (*e.g.*, an embedded version) and an ownership certificate (that includes a signed registration information of owner's identity, original, time of creation, etc.). Based on similar ideas, in [7], a cryptographic time-stamping service is used to certify the creation time of an object due to its security properties. In this model, to resolve ownership disputes the registration center needs to have access to original (or a faithful description of it) along with its registration time. The crux of this approach lies in the definition of ownership which refers not only to the original but for all works that are *similar* to the original. Therefore, a *similarity relation* needs to be defined and deployed to avoid multiple registrations of a particular work and to determine the owner of a given work. However, the applicability of this approach strongly depends on the availability of practical similarity measures that can reliably identify and distinguish various types of objects.

4. ESTABLISHING UNAMBIGUOUS OWNERSHIP

The watermarking based approaches that promise a solution to ownership problem, in fact, either solve it partially or raise new issues that need to be addressed. In this regard, most watermarking system designs were unable to capture the full scope of the problem whereas others required capabilities that are not yet available. Based on the discussion in the previous section, we can define the requirements for establishing unambiguous ownership on a cover-object as the following:

1. robustness;
2. low probability of false-positives;
3. non-invertibility; and
4. involvement of a trusted party.

The most trivial attack on the class of watermarking systems that intend to resolve rightful ownership is the removal of the watermark from the embedded-object. Therefore, the first requirement of an embedding/detection scheme is the robustness against all forms of malicious modification. Unfortunately, this is too strong an assumption to make in the presence of an intelligent and adaptive pirate, and the possibility of *unconditional* robustness is still an open question. However, *conditional* robustness can be achieved against a pirate who is restricted to a limited set of attacks. The second requirement refers to the ease with which a pirate can extract a fake watermark from an embedded-object which might potentially lead to an ownership deadlock or counterfeit ownership. The third requirement is to ensure that watermark embedding cannot be reversed so that a watermark cannot be subtracted from an object to produce a fake original. One way to cope with this is by making it difficult for the pirate to obtain a *meaningful* fake watermark. This is usually achieved by cryptographic means which essentially make watermark generation a one-way relation. It should be noted that if the probability of false-positives is very low, then the invertibility of an embedding scheme does not pose a significant risk. However in a more realistic setting, where the false-positive probability is relatively high,

non-invertibility complicates the task of the pirate. The last requirement is that a trusted party be in place to ensure that in the case of an ownership dispute the actual owner is involved in the dispute. Without satisfying these requirements, a watermarking system cannot be used to establish ownership and resolve ownership disputes.

In the context of ownership issues, the most overlooked design aspect of a watermarking system is the need for a trusted third party. *Essentially, this raises a much more fundamental question: Is watermarking necessary?* In other words, if indeed the involvement of a very capable and resourceful trusted third party, which has to ensure registration of all submitted cover-objects, store all registration records (including a *description* of the cover-object) and iterate a search over all entries prior to registration (of a new cover-object), is required do we really need watermarking as an essential component in a practical setup? This question has not been explicitly answered yet.

In [3], the authors argued that a solution to ownership problem requires the presence of a similarity relation that partitions the object space into (disjoint) equivalence classes so that only a cover-object and its variants are in the same class. However, an automated version of such an idealistic relation does not seem to be realistic. Therefore, one has to deploy a similarity relations that are not equivalence relations. (The use of similarity relations that impose partitioning of the object space into *intersecting* sets of objects implies that an object may be *similar* to many cover-objects thereby creating a confusion.) Two approaches have been considered to enable trusted party to create a registration record and to determine if and when a cover-object is registered. The first approach features a set of identifying characteristics to describe and register a cover-object. Hence, the ownership of an object, as well as the similarity of two objects, is established on the basis of object characteristics. However, as pointed by the authors a cover-object can be manipulated to yield different characteristics rendering the registration record useless. Alternatively, watermarking is considered as a solution in relation to its intended use. During registration the trusted party creates and embeds a watermark (as a part of the registration record along with the registration time and owner's identity) to the cover-object and returns the resulting embedded-object to the owner. Correspondingly, prior to registration of a new object the trusted party checks whether any of the registered watermarks is embedded in the object. Naturally, the success of this scheme depends on how well the requirements discussed in the previous section are met.

To rightfully establish ownership and resolve disputes in a practical setting, a trusted party has to rely on the existence of an automated procedure to obtain a *faithful description* of the object. It should be noted that due to storage, cost and security limitations, it is not practical to keep the originals. Therefore, the description should be a *digest* of the object which also enables easy search of *similar* objects. Furthermore, there should be some form of collision resistance so that it is very difficult to obtain an object given any *digest*. In the absence of such collision resistance, the system is prone to attacks by creating (or finding) and registering a *similar* of an unregistered object. These objectives can be achieved through the use of *robust perceptual hashes* which simply have a distinctive feature compared to cryptographic hashes. Ideally, a robust perceptual hash function maps *per-*

ceptually same or very similar inputs to a fixed data point in the hash space (robustness), whereas inputs sufficiently dissimilar are mapped to completely *unpredictable* data points (similar to cryptographic hash functions). *Therefore, with the availability of robust perceptual hash functions, ownership problem can be resolved through the involvement of a trusted third party, and watermarking is not needed.* The construction of robust perceptual hash functions is the subject of the ongoing research and current approaches to design of perceptual hash functions are not based on a formal definition as in the case of cryptographic hash functions. Therefore, only heuristic versions with poor collision properties are available. In other words, the security properties of perceptual hash functions are not well established and it is feasibly possible to find (or generate) two different inputs with the same perceptual hash.

It should also be noted that, from a practical standpoint, the notion of robustness of a perceptual hash function requires the flexibility that similar objects yield *very* similar hashes, within a tolerable limit, and not necessarily the same hash. The problem of representing a particular realization of input and all variants by a single data point (in the hash space) is equivalent to transforming the set of similar points into a perceptual space where the resulting transformed data points are arranged in a continuum which can be contained by a hyper-sphere. Unfortunately, such transforms are not readily available. Therefore, finding a collision of a perceptual hash functions simply requires finding two inputs that yield hashes that are *close* with respect to a distance measure (which is based on properties of the hash function). *As a consequence, despite their satisfactory robustness properties state of the art perceptual hash functions are not sufficient by themselves in solving the ownership problem, and some other mechanism is needed.* In this regard, the capabilities of watermarking techniques can be incorporated to overcome the deficiencies of perceptual hash functions.

4.1 Watermarking and Ownership Problem

Based on the above considerations, the key tasks of an ownership assertion system, that deploys the available perceptual hash functions along with watermarking techniques, can be identified as the following.

- At registration, the trusted party computes the perceptual hash of the cover-object and searches its records for hashes that are close to the obtained hash with respect to an appropriate distance measure.
- If no hashes were found, the trusted party creates (or facilitates the creation of) a watermark which will be embedded to the cover-object, records registration time and owner's identity, and provides owner with a signed ownership certificate as a proof of ownership.
- On the contrary, if the search yields hashes that are close to the computed hash, the trusted party checks whether the watermarks associated with the search results are embedded in the object. If none of the watermarks can be reliably detected registration continues as above, otherwise it fails.
- In the case of an ownership dispute, the trusted party searches its registration records to determine all the cover-objects (and the associated watermarks) which have perceptual hashes close to that of the object in

question. Then, the ownership is decided based on the extracted watermark.

The underlying assumption here is that watermark embedding and watermark removal attacks will cause minute changes on the perceptual hash of the cover-object since both the embedder and the attacker are distortion constrained and perceptual properties of the cover-object have to be preserved. This ensures that once a cover-object is registered it is not possible to claim counterfeit ownership on it at a later time. It should also be noted that in this setup ownership deadlocks cannot arise since registration time of a cover-object designates its owner.

The vulnerability of this system is mainly due to the possibility of collisions and the non-equivalence relation used to assess the similarity between two hashes. These may lead to two forms of attack:

1. The attacker may exploit this to claim counterfeit ownership on a group of *unregistered* objects. For this the attacker creates and registers many cover-objects which yield collisions with the designated (unregistered) objects. Once the number of registered objects reaches a certain number, depending on the false-positive probability of watermark detection scheme, the attacker will be able to detect one of the registered watermarks in the target objects.
2. In a similar manner, the attacker(s) may register many cover-objects to launch a *denial-of-service* attack on potential registrants. Since the trusted party has to search over all registered cover-objects prior to registration of a new object, and since the comparison of the hashes is not based on an equivalence relation the search may return many collisions. Trusted party has to ensure that the watermarks associated with the originals that yield a collision are not embedded in the object (to be registered). However, if the number of collisions is high, one of the registered watermarks can be detected and, therefore, registration of a legitimate cover-object might be unfairly denied.

Circumventing these attacks require that the probability of detecting a registered watermark in an unregistered object be very low. In essence, this refer to false-positive probability of watermark detection scheme.

5. REDUCING FALSE-POSITIVE PROBABILITY VIA MULTIPLE WATERMARK EMBEDDING

Let the cover-object C and embedded-object E be obtained from the set $\mathcal{C} = \mathbb{R}^n$, and the key K and watermark W be drawn from the set $\mathcal{K} = \mathcal{W} = \{0, 1\}^k$ where $k \leq n$. We define a proper subset $\mathcal{W}_{\hat{E}}$ of \mathcal{W} as

$$\mathcal{W}_{\hat{E}} = \{W \mid \mathcal{D}(\hat{E}, C, W, K) = \text{true} \quad \forall W \in \mathcal{W}\} \quad (3)$$

which consists of all the watermarks that can be *detected* in the object \hat{E} by detection rule \mathcal{D} . The set of watermarks that can be *reliably detected* in a given object by a specific detection technique, as in (3), will be referred to as *characteristic watermark set* of the object. It should be noted that *characteristic watermark set* of an object is strictly tied to specifics of the detection rule and it varies depending on

the use of original C or the key K in watermark detection and the criterion used to ensure reliable detection. For a particular cover-object $C \in \mathcal{C}$, $\mathcal{W}_C \subset \mathcal{W}$ represents the watermarks that are inherently present in C . Correspondingly, the low false-positive probability requirement of a watermarking system can be expressed as

$$|\mathcal{W}_C| \ll |\mathcal{W}| = 2^k. \quad (4)$$

If the probability of false-positives of an embedding/detection scheme is known to be p_{fp} , then $|\mathcal{W}_C| = p_{fp} \times 2^k$. Now, at the embedder the watermark W_E is embedded to C (under key K), yielding $E = \mathcal{E}(C, W_E, K)$. The *characteristic watermark set* of E is defined as

$$\mathcal{W}_E = \mathcal{W}_C + \{W_E\} \quad \text{and} \quad |\mathcal{W}_E| = |\mathcal{W}_C| + 1. \quad (5)$$

The watermark W_E is obtained through a one-way function due to non-invertibility requirement, *viz.* Section 4. This function takes as input the cover-object C , and it might further be keyed with a secret. For simplicity, we will assume that the watermarks are generated by a generic hash function, $h : \mathbb{R}^n \rightarrow \{0, 1\}^k$ with strong collision resistance properties.

In the general solution of the ownership problem, where a thrusted third party keeps digests and registration time of all cover-objects, the main concern is the collisions in perceptual hash functions. In this regard, first attack exploits this by registering a cover-object that yield a collision with an unregistered object and have the pirate's (registered) watermark embedded in it. Since, the difficulty of engineering a collision with state-of-the-art perceptual hash functions is rather low, the complexity of this attack is roughly in the order of $\frac{1}{p_{fp}}$. The increased number of collisions due to use of non-equivalence relations in searches is another vulnerability of this system which may cause a *denial of registration* condition. Let $h_p : \mathbb{R}^n \rightarrow \{0, 1\}^m$ be a perceptual hash function that maps an input cover-object to an m -bit sequence for $m < n$. For a perceptual hash function a collision is defined between two cover objects $C_1, C_2 \in \mathcal{C}$ such that $C_1 \neq C_2$ and $d(h_p(C_1), h_p(C_2)) \leq \epsilon$ where d is a distance (*closeness*) metric defined between two hash vectors and ϵ is a relatively small number. The proper measure of distance between two hash vectors depends on the construction of the particular perceptual hash function. However, for the general case we assume distance measure d is based on either the difference in certain number of least significant bits (LSB) or the hamming distance between the two vectors. The actual number of collisions, associated with a new cover-object submitted for registration, depends on the number of existing entries that have perceptual hash vectors *close* to the searched hash. However, the average number of collisions c can be obtained in terms of the total number of perceptual hashes η , already stored in the database of the thrusted party. Accordingly, when the distance is set to ϵ bits of difference in the LSB's of the compared hash vectors, $c = \frac{\eta}{2^m} \times 2^\epsilon$ and for a hamming distance of ϵ bits, $c = \frac{\eta}{2^m} \times \frac{m}{\epsilon}$. Assuming $\{C_1, C_2, \dots, C_c\}$ is the set of registered cover-objects that yielded collisions with a cover-object C , the probability of any of the watermarks associated with this set of cover-objects being in the *characteristic watermark set* of C is

$$\begin{aligned} &Pr(W_{C_1} \vee W_{C_2} \vee \dots \vee W_{C_c} \in \mathcal{W}_C) = \\ &Pr(W_{C_1} \in \mathcal{W}_C) + \dots + Pr(W_{C_c} \in \mathcal{W}_C) = c \times p_{fp}. \end{aligned} \quad (6)$$

Hence, the probability increases (almost) linearly with the number of registrations, and every $\frac{1}{c \times p_{fp}}$ other cover-object will be vulnerable to *denial of registration*.

Therefore, low values of p_{fp} is very crucial for successful operation of a watermarking based ownership system. However, when combined with pirate's ability to manipulate embedded-objects, achieving very low values of p_{fp} is not possible with the state-of-the-art watermarking techniques. Overcoming this dilemma requires alternate remedies that will effectively reduce the cardinality of the *characteristic watermark set* of a cover-object. To achieve this goal, we consider embedding multiple watermarks that are independent with each other. Essentially, the crux of multiple watermark embedding lies in the use of multiple one-way functions in generating watermarks (as opposed to only one) which makes brute search of fake watermarks more difficult. As a result, the ability to extract fake watermarks is not sufficient in itself to mount an attack unless all the extracted watermarks are interrelated by the predefined watermark generation criteria. For this, we assume a family of hash functions $\mathcal{H} = \{h_1, h_2, \dots, h_s\}$ to generate the set of watermarks $\{W_1, W_2, \dots, W_s\}$ from cover-object C such that $h_i(C_i) = W_i$. It should be noted that watermark generation can be achieved through various cryptographic constructions. For instance, the cover-object can be hashed at various hash-depths to yield many hash vectors to be used as watermarks, or the trusted party may split a secret obtained from the cover-object into many shares and designate some of the shares as watermarks.

Given an embedding/detection scheme and an embedded-object $E = \mathcal{E}(C, W)$, the associated probability of false-positives, p_{fp} , is a direct indicator of the difficulty of finding a random watermark in the *characteristic watermark set* of E , *i.e.*, $Pr(W^* \in \mathcal{W}_E) = p_{fp}$. In a similar manner, the probability of detecting a set of watermarks $\{W_1^*, \dots, W_s^*\}$ in E is

$$Pr(\{W_1^*, \dots, W_s^*\} \in \mathcal{W}_E) = Pr(W_1^* \in \mathcal{W}_E) \times \dots \times Pr(W_s^* \in \mathcal{W}_E) = p_{fp}^s. \quad (7)$$

Now, assume that this scheme is used to embed the set of watermarks $\{W_1, \dots, W_s\}$ in a cover-object C , *e.g.*, $E = \mathcal{E}(C, \{W_1, \dots, W_s\})$, rather than only W_1 . In this case, finding a false-positive requires extracting a set of *valid* watermarks $\{W_1^*, \dots, W_s^*\}$ from E , and correspondingly, the probability for this is p_{fp}^s . It should be noted that one does not have the freedom to determine *any* s random elements in \mathcal{W}_E since the watermark generation procedure requires the presence of a (fake) original C^* satisfying $h_i(C^*) = W_i^*$, $1 \leq i \leq s$. *Consequently, a linear increase in the number of embedded watermarks causes an exponential drop in the overall probability of false-positives.* However, in a watermarking scheme, the amount of distortion that can be introduced to a cover-object is perceptually constrained, and this is the main resource of the communication between embedder and detector. Since for a given cover-object embedding distortion is fixed, multiple watermark embedding requires dividing the permitted distortion among the watermarks to be embedded rather than using it up with a single watermark. Therefore, the question to be answered is whether multiple watermark embedding can improve the performance of watermark detector under distortion limited embedding scenario.

Although the actual realization of the watermark detec-

tion process varies with the embedding/detection technique, the detection process can be simply viewed as a procedure for statistically differentiating objects embedded a specific watermark from the rest of the objects. Alternatively, this problem can be formulated as a binary hypothesis test. For this, let the null hypothesis \mathcal{H}_0 be "*the object does not contain the specific watermark(s) in its characteristics watermark set*", and the alternative hypothesis \mathcal{H}_1 be "*the object contains the specific watermark(s) in its characteristics watermark set.*" Given an object O with unknown nature, the watermark detector tries to verify the presence of the watermark(s) by computing a test statistic d , which is essential in making a decision to accept (or reject) one of the two hypotheses. The performance of a watermark detector is evaluated by receiver operating characteristics (ROC) analysis. This is based on two measures, namely probability of detection p_d and probability of false-positives p_{fp} . The former refers to power of the detection test which is the probability of detecting the embedded watermark correctly, *i.e.*, $p_d = Pr(D(O, W) = true | \mathcal{H}_1)$, and the latter is the probability of detecting an un-embedded watermark, *i.e.*, $p_{fp} = Pr(D(O, W) = true | \mathcal{H}_0)$.

Consider the case of *single* watermark embedding where the permitted embedding distortion P_E is utilized by a single watermark. Let the corresponding test statistic be denoted by d_{one} , for a given watermark W_1 , and the two probabilities be denoted by p_d and p_{fp} . In the case of multiple watermark embedding, on the other hand, P_E is distributed among $\{W_1, \dots, W_s\}$ equally. Since the robust detection of a watermark in an embedded-object depends on the degree of embedding distortion and in multiple watermark embedding each watermark introduces a fraction of embedding distortion P_E to a cover-object, the corresponding test statistic d_{mul} under \mathcal{H}_1 for a given watermark, say W_1 , will be less reliable, compared to single watermark embedding. The s -fold decrease in P_E per watermark reflects also on detection and false-positive probabilities which are denoted in this case with p'_d and p'_{fp} . *However, one important difference is that when multiple watermarks are embedded a false-positive arises only if a set of s watermarks (seeded by the same cover-object) are detected and, therefore, it is not the false-positive probability due to detection of a given watermark.* That is the false-positive probability of multiple watermark embedding is obtained as $p_{fp}^{mul} = (p'_{fp})^s$. Whereas the detection probability of each watermark is defined as $p_d^{mul} = p'_d$. Correspondingly, multiple watermark embedding provides an advantage over single watermark embedding when

$$p_{fp}^{mul} = (p'_{fp})^s < p_{fp} \text{ for } p_d^{mul} = p_d. \quad (8)$$

This simply implies that if increasing p_d^{mul} to p_d induces a lesser increase in p'_{fp} with respect to p_{fp} , so that the exponential drop in $(p'_{fp})^s$ can compensate for that increase, then multiple watermark causes a reduction in the false-positive probability of watermarking scheme.

The most popular approach to watermark embedding and detection has been the additive watermarking technique described as

$$E_{one} = C + \alpha W_1 \quad (9)$$

where $E, C \in \mathbb{R}^n$, $W_1 \in \{-1, 1\}^n$, $\alpha \in \mathbb{R}$, and the distortion due to embedding is $P_E = \alpha^2$. Considering a set of independent watermarks $\{W_1, \dots, W_s\} \in \{-1, 1\}^{s \times n}$ to be

embedded in C , the embedding rule takes the form of

$$E_{mul} = C + \frac{\alpha}{\sqrt{s}}(W_1 + \dots + W_s) \quad (10)$$

where the total embedding distortion is $P_E = \alpha^2$.

In most cases additive watermarking techniques employ a correlation detector. In fact, assuming independent identically distributed (i.i.d.) zero mean Gaussian distributed cover-object samples and i.i.d. watermark samples, the optimal likelihood ratio detector is the correlation detector. Accordingly, the detector computes the detection statistic

$$d_{one} = \sum_{i=1}^{i=n} E_{one}[i] \times W_1[i] \quad (11)$$

and decides in favor of one of the hypotheses. To test the presence of W_1 , the two hypotheses are formulated as

$$\begin{aligned} \mathcal{H}_1 &: W_1 \in \mathcal{W}_O \\ \mathcal{H}_0 &: W_1 \notin \mathcal{W}_O \text{ (i.e., } O = C \text{ or } O = C + \alpha W^* \\ &\text{and } W^* \neq W_1) \end{aligned} \quad (12)$$

where O is an object whose type is in question. Watermark detector's decision is based on comparison of d_{one} to a threshold τ , which also designates the probabilities p_d and p_{fp} . Due to central limit theorem, test statistic d_{one} can be shown to be a Normal distributed random variable under both hypotheses. Hence, p_{fp} and p_d are computed as

$$p_{fp} = Q \left(\frac{\tau - E(d_{one}|\mathcal{H}_0)}{\sqrt{Var(d_{one}|\mathcal{H}_0)}} \right) \text{ and } p_d = Q \left(\frac{\tau - E(d_{one}|\mathcal{H}_1)}{\sqrt{Var(d_{one}|\mathcal{H}_1)}} \right) \quad (13)$$

where $Q(x)$ is the Gaussian error function defined as $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) dt$ and

$$\begin{aligned} E(d_{one}|\mathcal{H}_0) &= 0, \quad Var(d_{one}|\mathcal{H}_0) = N\sigma^2, \\ E(d_{one}|\mathcal{H}_1) &= N\alpha, \quad Var(d_{one}|\mathcal{H}_1) = N\sigma^2 \end{aligned} \quad (14)$$

where σ^2 is the variance of the cover-object.

One important aspect of multiple watermark embedding that needs to be mentioned is that the pirate should have a restricted access to watermark detector, as in the random oracle model. Otherwise, the pirate may exploit the linearity of the embedding scheme by enabling detection of the sum of watermarks $W_1 + \dots + W_s \notin \{-1, 1\}^n$ rather than detecting each watermark individually, thereby circumventing the security improvements offered by multiple watermark embedding. However, limiting the detector input to watermarks with sample values in $\{-1, 1\}$ renders summing watermarks ineffective, as the sum increases with the number of watermarks. In multiple watermark embedding, the detection statistic for each watermark is obtained as in (11) and the binary hypothesis testing of (12) has to be repeated for all watermarks. The probability of detection for each watermark p_d^{mul} and the false-positive probability of p_{fp}^{mul} of multiple watermark detection can be computed similar to (13) as

$$\begin{aligned} p_d^{mul} &= Q \left(\frac{\tau' - E(d_{mul}|\mathcal{H}_1)}{\sqrt{Var(d_{mul}|\mathcal{H}_1)}} \right) \text{ and } p_{fp}^{mul} = (p_{fp}')^s \text{ where} \\ p_{fp}' &= Q \left(\frac{\tau' - E(d_{mul}|\mathcal{H}_0)}{\sqrt{Var(d_{mul}|\mathcal{H}_0)}} \right) \end{aligned} \quad (15)$$

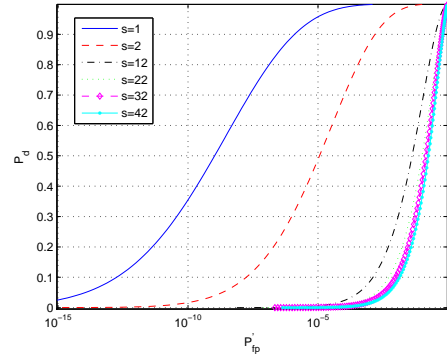


Figure 1: The increase in the probability of false-positive, p'_{fp} , for multiple watermark embedding.

and

$$\begin{aligned} E(d_{mul}|\mathcal{H}_0) &= 0, \quad Var(d_{mul}|\mathcal{H}_0) = N\sigma^2, \\ E(d_{mul}|\mathcal{H}_1) &= \frac{N\alpha}{\sqrt{s}}, \quad Var(d_{mul}|\mathcal{H}_1) = N\sigma^2. \end{aligned} \quad (16)$$

To compare the false-positive probability of multiple watermark embedding to single watermark embedding, the probabilities of detecting a watermark in both cases have to be equalized by properly adjusting the thresholds, so that $p_d = p_d^{mul}$. In order to detect each of the embedded watermarks as reliably as in the case of single watermark embedding, where for a given watermark embedding distortion is s times higher, the threshold τ' needs to be decreased to

$$\tau' = Q^{-1}(p_d) \times \sigma\sqrt{N} + \frac{N\alpha}{\sqrt{s}}. \quad (17)$$

Correspondingly, p_{fp} and p_{fp}^{mul} can be expressed as

$$p_{fp} = Q \left(\frac{\tau}{\sigma\sqrt{N}} \right) \text{ and } p_{fp}^{mul} = Q \left(\frac{\tau'}{\sigma\sqrt{N}} \right)^s \quad (18)$$

Figure 1 shows the the increase in the probability of false-positive detection p'_{fp} due to the reduction in the embedding distortion. The ROC curves for additive watermarking technique for varying numbers of s and n are displayed in Figure 2. The results show that with the use of multiple watermark embedding, at a fixed probability of watermark detection, false-positive probability reduces with the increase in the number of embedded watermarks.

The other important issue is the *robustness* of the multiple watermark embedding. Since the detection operation requires that each of the embedded watermarks be extracted *reliably*, the ability to deplete the integrity of only one watermark, while keeping the rest intact, will lead to a successful attack. For such an attack to be effective, the attacker should be able to introduce a controlled amount of distortion to render a specific watermark (or a few of them) undetectable. In other words, assuming embedded watermarks are independent, this requires the introduced distortion vector (due to attack) to have components only in the direction of certain watermarks. When the attacker has no *a-priori* information on the watermarks, it is reasonable to assume that the amount of distortion will be distributed over all watermarks almost equally. This will yield the same watermark to attack distortion ratio as in the case of single watermark

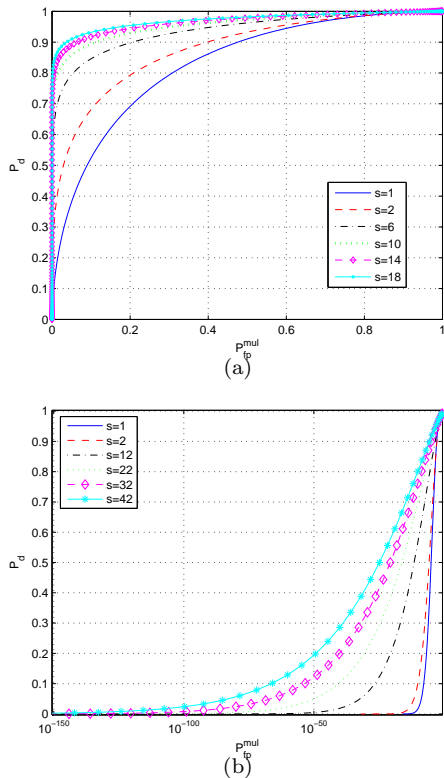


Figure 2: ROC curves corresponding to watermark detection under $\mathcal{H}_1 : O = E + \frac{\alpha}{\sqrt{s}}(W_1 + \dots + W_s)$ and $\mathcal{H}_0 : O = C$ for varying s , $\sigma = 100$, $\alpha = 6$ and (a) $N = 500$ (b) $N = 5000$.

embedding, and therefore, multiple watermark embedding will not cause a reduction in the robustness. For example, if the attack distortion is independent of the embedded watermarks, the results of additive embedding scenario discussed above will still be valid except for the modifications in (14) and (16) to include the statistics of attack distortion. These new terms may effectively be absorbed into the statistics of the cover-object in both cases yielding the same formulation. Hence, the reduction in the false-positive rate will not be accompanied by a reduction in the robustness.

6. CONCLUSIONS

In this paper, we have revisited the ownership problem and assessed the role of watermarking in devising a solution to this problem. With this intent, we examined watermarking based approaches to ownership problem and identified their strengths and limitations. Based on these deductions, we determined the requirements of a practical watermarking based ownership assertion system. The successful operation of this system relies on robustness properties and the false-positive probability of the underlying watermark embedding/detection scheme. To make attacks due to high false-positive rates more difficult, we considered and analyzed embedding multiple watermarks rather than a single one, under constrained embedding distortion. In this approach, each watermark is generated from the given cover-object using a different one-way transformation. We em-

ployed multiple watermarking technique in conjunction with additive watermarking technique. The results show that the false-positive probability indeed decreases with the number of embedded watermarks. The robustness of the multiple watermark embedding scheme will be further analyzed.

7. REFERENCES

- [1] I. J. Cox, F. Kilian, F. T. Leighton, and T. G. Shamoan, "Secure spread spectrum watermarking for multimedia," *IEEE Transaction on Image Processing*, vol. 6, no. 12, pp. 1673–1687, 1997.
- [2] S. Craver, N. Memon, B. Yeo, and M. Yeung, "Can invisible watermarks resolve rightful ownerships," in *Technical Report RC 20509*. 1997, IBM Research Institute.
- [3] A. Adelsbach, B. Pfitzmann, and A. R. Sadeghi, "Proving ownership of digital content," in *Proc. of IHW'99, Lecture Notes in Computer Science*. 2000, vol. 1768, pp. 126–141, Springer-Verlag.
- [4] S. Katzenbeisser and H. Veith, "Securing symmetric watermarking schemes against protocol attacks," in *Proc. of SPIE: Security and Watermarking of Multimedia Contents*, 2002, vol. 4675, pp. 260–268.
- [5] A. Adelsbach, S. Katzenbeisser, and H. Veith, "Watermarking schemes provably secure against copy and ambiguity attacks," in *Proc. of ACM CCS-10 Workshop on Digital Rights Management*, 2003.
- [6] A. Adelsbach, S. Katzenbeisser, and A. Sadegi, "On the insecurity of non-invertible watermarking schemes for dispute resolving," in *Proc. of IWDW*, 2003.
- [7] A. Adelsbach and A. R. Sadeghi, "Advanced techniques for dispute resolving and authorship proofs on digital works," in *Proc. of SPIE: Security and Watermarking of Multimedia Contents V*, 2003, vol. 5020.
- [8] A. Adelsbach, M. Rohe, and A. Sadeghi, "Security engineering for zero-knowledge watermark detection," in *Proc. of WIAMIS'05*, 2005.
- [9] A. Adelsbach, M. Rohe, and A. Sadeghi, "Overcoming the obstacles of zero-knowledge watermark detection," in *Proc. of ACM Multimedia and Security Workshop*, 2004, pp. 46–55.
- [10] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "The watermark copy attack," in *Proc. of SPIE: Security and Watermarking of Multimedia Contents II*, 2000, vol. 3971, pp. 371–380.
- [11] S. Craver N. Memon B. Yeo and M. Yeung, "Resolving rightful ownership with invisible watermarking techniques: Limitation, attacks, and implications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 4, pp. 573–586, 1998.
- [12] L. Qiao and K. Nahrstedt, "Watermarking methods for mpeg encoded video: Towards resolving rightful ownership," in *Proc. of ICMCS*, 1998, vol. 9, pp. 194–210.
- [13] W. Zeng and B. Liu, "On resolving rightful ownerships of digital images by invisible watermarks," in *Proc. of ICIP*. 1997, pp. 552–555, IEEE.
- [14] R. B. Wolfgang and E. Delp, "A watermarking technique for digital imagery: Further studies," in *Proc. of SPIE: Voice, Video and Data Communications*, 1997, pp. 297–308.
- [15] M. Ramkumar and A. N. Akansu, "Image watermarks and counterfeit attacks: Some problems and solutions," in *Proc. of Content Security and Data Hiding in Digital Media*, 1999.
- [16] M. Ramkumar and A. N. Akansu, "A robust protocol for proving ownership of multimedia content," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 469–478, 2004.
- [17] Q. Li and E. Chang, "On the possibility of non-invertible watermark schemes," in *Proc. of IHW'04, Lecture Notes in Computer Science*. 2004, vol. 3200, pp. 13–24, Springer-Verlag.