TDDC03 Projects, Spring 2004

Classifications of Attacks on Watermarking

Shanai Ardi Haiyan Jiao

Supervisor: Jacob Löfvenberg

Classifications of Attacks on Watermarking

Shanai Ardi
Department of computer and information science Linköping University
Shaar368@student.liu.se

Haiyan Jiao
Department of computer and information science Linköping University
Haiji374@student.liu.se

Abstract

Intellectual property protection is one of the most important features in the new digital world.
Digital watermarking embeds marks in digital content for the purpose of communicating copyright and ownership information. Strength and cleverness of attacks in watermarking needs precise efforts to keep the watermarked data tamper resistant. Information of classes in watermarking attacks helps to find the correct attempts to defeat this technique against malicious efforts. In this paper, introduction of some existing classifications of attacks on watermarking is given, and also references to find details about the purpose of these classifications. These classifications are analyzed and compared and a more generalized classification is presented.

1. Introduction

People have begun to study ways of embedding hidden marks and serial numbers in audio and video, to avoid illegal copying because digital media could be copied easily. One kind of such marks is watermark, which is "a digital code irremovably, robustly, and imperceptibly embedded in the host data and typically contains information about origin, status, and/or destination of the data". [6]
A good watermarking should be robust not only against standard data manipulations and format conversions, but also against attacks, which could be performed intentionally or unintentionally.
"Watermarking systems utilized in copy protection or data authentication schemes are especially susceptible to intentional attacks, which are usually done by more competent people with more knowledge of watermarking systems and more resources to make the attack, while the unintentional attacks usually come from common signal processing operations done by legitimate users of the watermarked materials." [5]
But after all attacking the watermark is not so difficult because as it is stated by [3] blind use of simple manipulations and studying the methods can show the week points since there is no standard and general-purpose benchmark. In this paper, an introductory explanation of the principles that apply to the watermarking techniques is given first; independent of the actual application, and then it gives detailed insight into attacks classification, and example for those classes.
Finally, it presents the conclusion and summary of this paper. There are many papers, which have focused on attacks and try to classify them. Five separate classifications are presented in section 4 and analyzed in section 5 and 6.

2. Watermarking Principles

"A digital watermark is a piece of information that is hidden directly in media content, in such a way that it is imperceptible to a human observer, but easily detected by a computer."[6]
In the other words a mark is embedded in original data and makes it difficult to remove the mark without degrading the original data.
There are three main issues in the design of a watermarking system:
1) The watermark data to be added to the host signal. Typically, watermark signal depends on key and watermark information.
2) The embedding method that incorporates the watermark signal to the host data.
3) Extraction method that recovers the host data from the watermarked data.
Figure 1 and 2 show the generic watermarking scheme. This scheme has been mentioned in [6] but there are schemes focused on more details on the other resources according to the points that are important for the writer.

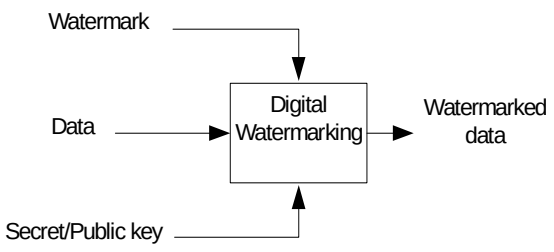

Figure-1
Generic digital watermarking scheme

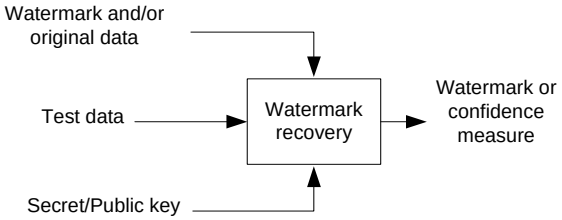

Figure-2
Generic watermark recovery scheme

In [9] the authors have mentioned another model of watermarking system according to a communication formulation. Its block diagram consists of three main parts: message embedding, attack channel and message extraction.
There are requirements, which should apply in watermarking systems, such as security, imperceptibility permanence and etc.
As can be seen in the scheme, in order to ensure the requirements of imperceptibility and security, watermark systems usually use a perceptibility criterion of some sort and one or more cryptographically secure keys, and the watermark information is usually redundantly distributed over many samples to ensure robustness.
The watermark signals that are to be added to the host signal typically depends on key and watermark information, and sometimes, also the host data. The key could be secret or public key to enforce the security of watermarking.
The watermark signal should be embedded into the host data to get watermarked data using some specific method. Correspondingly, there should be an extraction method to recover the watermark information from the signal mixture by the key.

3. Attacks on watermarking

Generally the goal of attacking is to remove or destroy the watermark while preserving the quality of the host data.
If x, n, y and y´ represent the original data to be covered, the noise-like watermark added to the original data, stego-data or watermarked data, and attacked Stego-data respectively; then the attack process can be shown with mathematical notation as below:
$y = x + n$; to preserve the data quality after attack it is necessary to have: $y' \cong x$ (almost equal).
As mentioned before a lack of systematic benchmarking of existing methods creates problems for watermarking technology suppliers and makes it is easier to attack watermarked data.
Enough information on attacks and attacks classifications is important in order to create tools for preventing the attacks.

So the next sections will focus on five different classifications, which have been done on the existing attacks.

4. Attacks classifications

Five classifications on watermarking attacks have been introduced. The descriptions are taken from each respective paper. These different classes are referred by the name of authors of the article that these classifications have been introduced.

Craver et al.

In this article [1], the authors state that the watermarking is robust if the added label or traces of the label are detectable in watermarked data. The robustness feature separates watermarking from other forms of data hiding. According to the paper the attacker wants to eliminate or degrade the effectiveness of owner's mark inserted for protection in order to control the watermarked content. Craver et al. have recognized four classes of attacks on watermarking schemes (only one of which requires a watermark's presence to be removed or diminished). Robustness is necessary but not sufficient to guarantee security. The four classes are:

Robustness attacks: These attacks aim to diminish or remove the presence of a watermark in marked data without harming the image. Generally such attacks modify the pixel values of the image. An example is UnZign [7], which applies noise to a watermarked image.

Presentation attacks: This attack does not necessarily remove the watermark from the content; instead it manipulates the content so the detector cannot find it. A good example is the Mosaic attack, which posses the initially remarkable property that the marks from an image can be removed and still have it rendered exactly the same, pixel for pixel, as the marked image by a standard browser. This attack consists of chopping an image up into a number of smaller sub images, which are embedded one after another in a web page. For more information, see [2].

Interpretation attacks: In some watermarking schemes, the mark's detected presence can cause multiple interpretations of the derived analytic or scientific data, and an attacker can engineer a situation that neutralize the strength of any evidence of ownership presented. These attacks seek to forge invalid or multiple interpretations from watermark evidence. For instance, and attacker can attempt to make another watermark in the same watermarked image with strength equal to

the first one in order to make it difficult to say which mark has been made first. An example is collusion attack described in [8].

Legal attacks: This attack may involve existing and future legislation on copyright laws and digital information ownership, the different interpretations of the laws in various jurisdictions, etc. An attacker can cast doubt on the watermarking scheme in the courtroom in order to cause problem to decide who the owner is. No examples were found for this class.

Fabien et al.

Fabien et al. [2], refer to the classification in [1] but they think that the separation between these classes is not always very clear though; for instance, StirMark[1] both diminishes the watermark and distort the content to fool the detector.
They develop a general attack based on simple signal processing, plus specialized techniques for some particular schemes, and show that even if a copyright marking system is robust against signal processing; bad engineering can provide other avenues of attacks.
So they add a new group of attacks to [1], called basic attacks. Also they introduce an attack on echo hiding as an example of the robustness attacks class.
However, according to the authors, the mosaic attack can be a new class in addition to [1].

Voloshynoyskiy et al.

According to Voloshynovskiy et al. [3], watermark attacks have three classes while ignoring the cryptographic and system-based attacks (e.g. Oracle, counterfeit original, averaging). The classes are:
Geometric attacks: (desynchronization) e.g. random local distortion.
Signal processing attacks: e.g. noise addition, dithering, and stochastic attacks
Specialized attacks: based of knowledge of the method

In another paper [9] a new way of the classification on attacks in watermarking comes up.

- Attacks concerning the statistics of the image. The main idea of these attacks is to perform watermark estimation and then remodulate the watermark by means of subtracting the estimated watermark from the stego data with some constant strength factor.

- Attacks to linear watermarking methods.

As mentioned before this article models the watermarking system with three main parts and defines the watermark attacks based on the weak points of linear methods:

- The watermark removal based on denoising/compression that uses the assumption of key-independent watermark extraction aimed at reducing the watermark redundancy.
- Perceptual remodulation of the watermark aimed at creating the least favorable statistics for the decoder designed for AWGN[2].

Hartung et al.

Hartung et al. [4], suggest another classification by reviewing proposed attacks on watermarking and they consider only attacks that do not significantly impair the perceived fidelity of the host data.
They distinguish among the following groups:

Simple attacks: (other possible names include "waveform attacks" and noise attacks). These are conceptually simple attacks that attempt to impair the embedded watermark by manipulation of the whole watermarked data (host data plus watermark) without any attempt to identify and isolate the watermark, like linear and general non-linear filtering.

Detection-disabling attacks: (other possible names can be synchronization attacks and Jitter attack). The main characteristics of this category is that an attacker does not attempt to remove the watermark from the watermarked data, but to remove the synchronization of the watermark so that it cannot be detected properly by the watermark detector. So the watermark itself may (and usually) still be physically present in the data. One example is Mosaic attack.

Removal Attacks: attacks, which aim to remove or seriously degrade the watermark, embedded in the watermarked data so that the detector can no longer positively detect it. It is further divided into "simple" and "analysis" attacks to show the different strategies taken to reach this common

[1] Stirmark is a benchmarking tool for basic robustness testing of image watermarking algorithm. It applies a minor unnoticeable geometric distortion: the image is slightly stretched, sheared, shifted, bent and rotated by an unnoticeable random amount. This method has been developed by Fabien Petitcolas and has gained large interest from the watermarking community.

[2] AWGN: **A**dditive **W**hite **G**aussian **N**oise

goal. Examples of these sub categories are lossy compression and non-linear filtering respectively.

Ambiguity Attacks: An attacker tries to embed another watermark into a watermarked data and thus making it difficult to determine the first embedded watermark.
Authors believe that the transitions between the groups are sometimes fuzzy, and some attacks do not clearly belong to one single group. Cropping for example can be regarded as either simple attacks or detection-disabling attacks. There is a table of classification in this article for proposed attacks that could be useful.

Setyawan

Setyawan [5], separates the components of watermark system into four main components that can be attacked and classifies the attacks according to the target component of attack. These components are: Data (host data to be marked) Watermark embedding algorithm, watermarked data and Watermark detection algorithm.
According to this assumption he defines two general classes for the attacks:
Class A: Attacks applied to the watermarked data
He defines two subclasses for class A.

- Attacks applied to the embedded watermark, which can be in three different categories. These categories are similar to those in [4], but here the “simple” attacks are together with “removal” attacks because the differences are slight.
- Attacks applied to the data portion of the watermarked data (content tampering).

These attacks are aimed to modify or tamper with the data in which the watermark is embedded, without destroying the watermark itself. This kind of attacks might for example be performed to discredit an institution by tampering with material publishing an allegedly compromising picture of somebody.
Class B: Attacks applied to the other components of the watermarking systems:
In this class an attacker attempts to defeat the watermarking system by attacking components other than the data. An attacker could attack the software or the hardware components. Hacking and cracking attacks and hardware tampering are from this class. To clarify the classes and attacks this article

5. Evaluation

After mentioning each classification above, they are going to be analyzed one by one in this section.

[1] Craver et al. addressed the classification of attacks in current watermarking schemes according to four-stage life cycle: representation, detection, judgment, and final decision. There are several advantages of this classification:

- They focused on the stages of life cycle in their classification and generally, attackers want to eliminate or degrade the effectiveness of owner's mark inserted for protection to control the watermarked content. So this classification can cover most of the attacks on watermarked data.
- Legal attacks, they are important and reality in the human social, but seldom mentioned in other classifications.

But there are also some disadvantages:

- None of these classes concern the attacks on the watermarking system.

There might be attacks satisfying the requirement of more than one class, like StirMark.

[2] Fabien et al. agree with classification of Craver et al. and they only add one more class, basic attacks on watermarked data, this helps to classify more attacks, but still has the disadvantages of the [1].

[3] Voloshynovskiy et al. suggest two kinds of classifications, one, which ignores the cryptographic attacks and system-based attacks, specifies three main classes: geometric, signal processing and specialized attacks. The other classification looks generally to two classes: stochastic attacks and attacks to weak point of linear methods.
Advantages for this classification:

- It nearly covers all the classes of attacks, especially "specialized" could cover all the other attacks rather than geometric and signal processing.
- They classify the attacks from another point of view, totally different with the others. So it suits the situation when the others do not work.

Disadvantages:

- The classes are relatively broad compared with [1], so in some cases, it could not figure out exactly what the attacks are.
- There might be attacks satisfying the requirement of both geometric and signal processing.

[4] Hartung et al. classify the attacks by reviewing proposed attacks on watermarking, considering only attacks that do not significantly impair the perceived fidelity of the host data. They list conceivable attacks on spread spectrum watermarks, but they think that the transitions between the groups are sometimes fuzzy, and that some attacks do not clearly belong to one group.

Advantages:

- According to their considerations and assumptions, this classification can cover almost all attacks on watermarked data.
- It is precise to separate simple and removal attacks, even if there are some attacks belong to both of them.

Disadvantages:

- None of these classes concern the attacks on watermarking system.
- They did not consider the social elements, like legal attacks in [1].

[5] Setyawan's article is the only one that has very clear classification for the watermarking systems. This article separates the components of watermark system into four main components that can be attacked and classifies the attacks according to the target component of attack. Advantages:

- This classification is general and by comparing the other classifications with it, a conclusion of existing classifications could be drawn out.
- Can cover both the attacks on watermarked data and watermarking system.
- No overlap between class A and class B.

Disadvantages:

- It does not concern the social factors, like legal attacks in [1].

After all these points, it comes to the conclusion that Setyawan's classification is more general than the others.

6. Results

In last section the articles that mentioned before was analyzed, and here they are classified according to the components of watermarking system that they focus in the classifications and by comparing them to [5]. This can help us to have a final general idea of these classes.

The classification done by Craver et al. and completed by Fabien et al. introduces the classes of attacks that are applied to the watermarked data. All of these classes are referring to attacks that try to eliminate the efficiency of watermark by removing the watermark (Robustness attacks), manipulating watermark and make it difficult to be detected (presentation attacks) or adding a new mark in order to cause an ownership deadlock (interpretation attacks). According to classes in [5] these classes belongs to class A. For example robustness attacks is one kind of simple-removal attacks category, presentation attacks can be in jitter category and interpretation attacks fits to analyze-removal attacks category.

Voloshynivskiy et al. have mentioned classes of attacks to watermarked data and watermark embedding algorithm. Classes like geometric, signal processing and stochastic attacks are applied to watermarked data and by reviewing them it becomes clear that geometric attacks is a kind of jitter attacks of class A category in [5], signal processing attacks is in simple-removal category and stochastic attacks fits to analyze-removal attacks category.

Specialized attacks and attacks based on the weak points of the method are applied to the components of the system other than data and they can be in class B of [5].

Hartung et al. introduces four classes that all of them are about attacks to watermarked data. These classes are matched with attacks applied to the embedded watermark data that its attacks are named class A-1 in [5] but they have only different names.

For example detection-disabling attacks can be in Jitter attacks category of class A.

Finally, a general classification is pointed out according to these references.

It is shown by the figure 3. As it is clear this classification follows the classification done by Setyawan [5]. It separates the attacks according to their target using the components of watermarked system defined in [5] and adds the legal attacks class introduced by Craver et al. [1].

By this consideration this classification can cover attacks to the data, watermarked data and the components of the watermarking algorithm as well as attacks on social features of the watermarking.

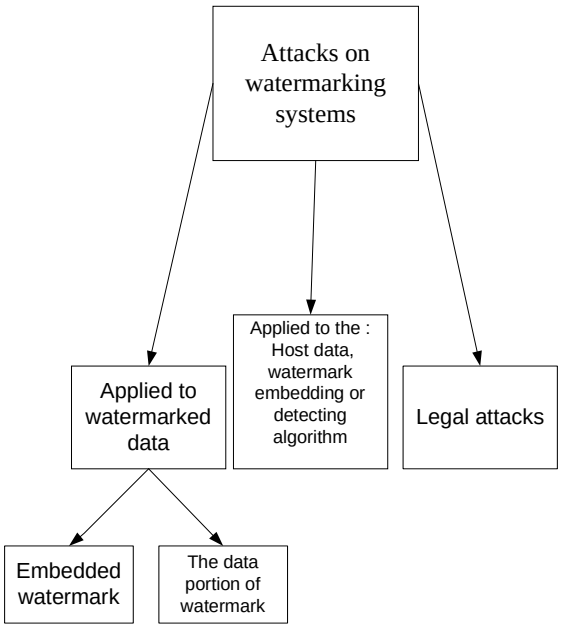

Figure-3
General classification of watermark attacks

7. Conclusions

Because both the watermarking technology and the attacks on watermarks will evolve, careful overall

system design under realistic expectations is crucial for successful applications. In this paper several classes of attacks have been outlined and explained. Different classifications of watermarking systems were introduced with some examples of each class/category.
As shown, watermarking systems are susceptible to many kinds of attacks and the information about different classifications can help to researchers to find useful tools for providers of watermarking systems to have more reliable techniques.

6. References

[1] S.Craver, B. -L. Yeo, and M. Yeung, "Technical trial and legal tribulations." *Communications of the A.C.M.,* Vol 41, no. 7, pp. 44-45, Jul 1998.

[2] Fabien A. P. Petitcolas, Ross J. Anderson and Markus G. Kuhn. "Information hiding – a Survey". *Proceeding of the I.E.E.E.,* 87(7): 1067-1078, Jul 1999.

[3] Sviatoslav Voloshynovskiy, Shelby Pereira and Thierry Pun, Watermark attacks, In *Erlangen Watermarking Workshop*, Erlangen, Germany, 5-6 October 1999. (invited presentation)

[4] F. Hartung, J. Su, and B. Girod, "Spread Spectrum watermarking: Malicious attacks and counterattacks" *Proc. SPIE Security and watermarking of Multimedia Contents 99*, San San Jose, CA, Jan. 1999.

[5] I. Setyawan, "Attacks on Watermarking Systems", *Technical Report*, Information and Communication Theory Group, TU Delft, 2000

[6] F. Hartung and M. Kutter, "Multimedia watermarking techniques*", Proc. IEEE,* vol. 87, no. 7, pp. 1079-1107, Jul. 1999.

[7] M. Kuhn, Web page on UnZign (see www.altern.org/watermark/index.html www.stealthencrypt.com/watermk.html)

[8] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia" *Tech. Rep. 95-10*, NEC Research Institute, Princeton, N.J., 1995.

[9] S. Voloshynovskiy, S. Pereira, A. Herrigel, N.Baumgartner, T. Pun, "Generalized watermarking attack based on watermark estimation and perceptual remodulation". *Proceedings of SPIE: Security and Watermarking of Multimedia Content II*, San Jose, CA, USA, Jan. 2000. SPIE.

[10] N. F. Johnson, and S. Jajodia, "Steganalysis of Images created using current steganography software", in *D. Aucsmith (Ed.): Information Hiding, LNCS 1525*, pp. 32-47. Springer-Verlag Berlin Heidelberg 1998.

[11] http://www.watermarkingworld.org