A Review on Attacks, Problems and Weaknesses of Digital Watermarking and the Pixel Reallocation Attack

K. F. Tsang, O. C. Au
Department of Electrical and Electronic Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong

ABSTRACT

Watermark attacks are first categorized and explained with examples in this paper. We then propose a new image watermark attack called "Pixel Reallocation Attack". The proposed attack is a hybrid approach, which aims to decorrelate the embedded watermark with the original watermark. Since many watermarking detections are by correlating the testing image with the target watermark, it will not work once we decorrelate the embedded watermark. For example, the geometrical transformation attacks desynchronize the correlation detector with the testing image leading to detection failure. However, by inserting a template or grid into the watermarked image can make inverse transformation possible and the watermark can be retrieved. If we apply transformations to every single pixel locally, independently and randomly, inverse transformation will not be possible and the attack will be successful since the embedded watermark is not correlated with the original watermark. Experiment shows that single technique approach needs a larger distortion to the image in order to attack the image successfully. We also tested our attack with commercial watermarking software. It cannot detect the watermark after we applied the proposed hybrid attack to the watermarked image.

Keywords: Watermark, Attacks, Pixel reallocation, Robustness, Copyrights protection

1. INTRODUCTION

In the development of digital watermark, many researchers focus on applying digital watermarking techniques to copyrights protection, proof of ownership, etc. Generally, author, copyright holder and/or the customer's information are embedded in the digital media as a watermark. Some applications also store access control parameters, like the number of allowable time of access, in the watermark.

Robustness is a key requirement for these applications. That is, the watermark embedded in the data must be recoverable despite intentional or non-intentional modification of the watermarked data – a requirement generally very difficult to meet. For example, the watermark should be robust against such signal-processing operations as filtering, requantization, cropping, compression, and so on. Clearly, there is a need of identifying all the possible attacks, which the embedded watermark will face during its transmission, detection and/or extraction, both intentional and non-intentional, before designing a robust watermarking scheme.

However, watermark designer has only limited knowledge about the predicted attacks and even less for the unforeseeable new attacks during the design. It is very difficult to ensure the watermark will survive under these unforeseeable attacks. This is especially true for those watermark applications like robust watermarking for copyrights protection or access control (e.g. SDMI), which requires highest robustness and is subjected to all kind of attacks in various conditions.

In this paper, we study different types of attacks to digital watermarking, based on a generalized digital watermarking model, which fits to many common digital watermarking techniques. In Section 2, we describe the said digital watermarking model. Different types of attacks are classified in Section 3. Non-Intentional and Intentional attacks are discussed in detail in Section 4 and 5 respectively. We then propose a new watermark attack in section 6 to show that simple but efficient attack can be developed easily. Experimental results are shown in Section 7. We conclude in Section 8 and future directions are discussed.

2. DIGITAL WATERMARKING MODEL

In recent years, most of the watermarking techniques tend to model the watermark as information that transmit in a communication system and the host signal or media as channel noise. Like the one proposed by Cox et al. [2], the watermark string (a PN sequence with length N) is added to the N largest DCT coefficients of the original image. Many techniques proposed later were based on this approach. Another common watermarking techniques is by generating a noise like PN sequence with a secret key and adding to the temporal domain of the host signal. For example, Swanson et al. [11] proposed to mask the noise like PN sequence with the Human Auditory System before adding to the original audio signal, to ensure the imperceptibility of the watermarked audio signal.

Mathematically, most of the watermarking techniques, including the above two techniques, can be modeled as follow [3]: Given an original media M, the watermarked media M' is formed by $M' = M + f(M,W)$ such that people would not find any difference between the original media M and the watermarked media M'. Function f is the most critical part of a watermarking algorithm and the robustness of the watermark depends greatly on the design of f. Common detection algorithm is to correlate the watermark with the watermarked media. If W is chosen at random, the correlation between M and W is very small, as the random positive and negative terms tend to cancel themselves out, especially for long sequences. However, in computing $W \bullet W$ all of the terms are positive, adding up to a large value compared to $M \bullet W$. For this reason, the product $M' \bullet W = M \bullet W + W \bullet W$ can estimate $W \bullet W$ accurately. As a result, the embedded watermark can be retrieved without the use of the original media M.

3. CATEGORIZATION OF DIGITAL WATERMARK ATTACKS

There are two main categories of Watermark Attacks: Non-Intentional Attacks and Intentional Attacks.

Non-Intentional Attacks
Non-Intentional Attacks refer to those attacks, which are not imposed by an attacker. This class of attack is introduced mainly during the transmission of the media. For example, the embedded watermark will be lost if the image is compressed with JPEG codec, printed to paper (D/A conversion); the audio is recorded to a analog tape; noise addition when using wireless transmission, etc. Any watermarking techniques must be robust to this class of attack, since the media must be distributed from the content owner to the end user with some means of transmission.

Intentional Attacks
Intentional Attacks can be further sub-categorized to [5]: Detection Disabling Attacks, Removal Attacks and Ambiguity Attacks.

Detection Disabling
Detection Disabling Attack aims at making the correlation detector, which is a common detector, fail to detect or extract the watermarked info. This class of attacks also know as Geometric Attack, because many attacks in this class are achieved with geometric distortion like zooming, shifting in spatial or temporal domain, rotation, shearing, cropping, sub-sampling, removal or insertion of pixels, or any other kinds of geometric transformations. Since this class of attack does not remove the watermark info from the media, by using more sophisticate watermarking techniques may be able to recover the watermark from the attacked media, like [10], [7].

Removal Attacks
Removal Attack aims at destroying or removing the watermark or part of the watermark from the watermarked signal, such that the correlation detector cannot extract the watermark or the resultant score of the detection is lower than the threshold. For invertible watermarking techniques, attacker can look into the watermark insertion or extraction algorithm in the present of the inserter or the detector. Once the attacker figures out how the watermarking process is, a simple inversing processing can completely remove the watermark from the watermarked signal. Attacks like denoising, certain non-linear filtering operations, compression/decompression, statistical averaging, bit-by-bit removal, etc., can also diminishing the watermark, such that the output score of the detector is lower than the threshold.

Ambiguity Attacks

Ambiguity Attack refers to those attacks to the watermarking scheme or system, instead of the watermarking algorithm/techniques. Most of these attacks origin from the question "What is a watermark?". One can watermark a watermarked signal again with the same or different watermarking technique, such that it is unclear which is the first and authoritative watermark of the owner. Some techniques proposed to make a fake original from the watermarked signal, with the presence of the detector. Attackers can also easily to false the watermark detector in some access control system, which use watermark to store the access control information [3]. Sometimes we refer these kinds of attack to "deadlock problem", "IBM attacks" [4] and "confusion attacks". In the case of proving the ownership of the media, an even more simple attack which can apply is to pick some bits randomly from the watermarked or non-watermarked signal, apply an arbitrary function to it and claim that it is the watermark of the attacker. This problem exits because there is no answer to the question, "What is a digital watermark?".

4. NON-INTENTIONAL ATTACKS

In this section, we summarize common non-intentional attacks, which mostly are those signal processing imposed on the watermarked media during distribution.

A) Attacks induced by transmission

Noise is always introduced to the media when transmit through an analogue channel. Examples are analogue transmission of TV/Radio signal, printing an watermarked image on paper, etc. However, noise is not a serious attack to most of the digital watermarking techniques, unless the noise is large compared to the host signal. Assume $M' = M + W + n$, where n is the noise. $M' \bullet W = M \bullet W + W \bullet W + n \bullet W$. Since $W \bullet W$ is much larger than $M \bullet W$ and $n \bullet W$, the detection of the watermark is still success. Geometric transformation is another common attack induced during transmission. Transformation like shift, rotation, scaling, shearing, etc., would lead to the desynchronization of the detector with the embedded watermark. This is true for both cases of which the watermark is inserted to spatial domain or the frequency domain. This kind of attack is also used commonly by attacker intentionally, since one or two degrees of rotation is not detectable for human eye. However, this little rotation could make the watermark detection fail.

B) Digital Compression

Media files are always transmitted from distributors to end-users in a compressed format, since this kind of files are usually very large in size (especially video and audio files). However, most of these compressions are lossy compression. In such a situation, the embedded watermark will be removed or partly removed, especially those are inserted to the high frequency portion of the signal, since common lossy compression removes the high frequency part of the original signal. Cox [2] proposed to embed the watermark to the perceptual significant part of the media, in order to make the watermark to be robust to compression and other filtering attacks (since removing such a watermark will also degrade the host media). On the other hand, this introduced another challenge for the watermarking design, which is how to maintain the invisibility or inaudibility of the watermark.

5. INTENTIONAL ATTACKS

Examples of intentional attacks, are discussed in this section.

A) Statistical Averaging/Block Averaging

Statistical Averaging is a very common removal attacks to digital watermark, especially for those watermarking techniques in which $U=f(M,W)$ does not depend significantly on the host media M [3]. This attack can be easily applied on audio signal, since most of the audio watermarking techniques, like in [11], insert watermark into individual frames. Averaging can then be applied to the frames of the audio signal. One simple solution to this attack is to use at least two different watermarks randomly. More advance approach like the one in [9] uses different frame size, different watermark and even different watermarking technique for each individual frames. This introduces more parameters which are unknown to the attacker and thus increases the robustness of the watermarking scheme.

B) Signal Processing Attack

Signal Processing Attack is a kind of removal attacks which use signal processing technique to remove the embedded watermark from the host media.

Filtering, using lowpass, bandpass and highpass filters, is a common signal processing attack. Since the watermark of many watermarking techniques is noise-like, filtering can effectively remove this noise-like watermark. However, this process will also degrade the host signal. The amount of degradation depends on how well the watermarking techniques embed the watermarks. For example, if the watermark is embedded in perceptually significant regions of the host signal (like the one proposed by Cox [2]), the watermark can only be removed with the host signal suffering from serious degradation. On the down side, embedding watermark in perceptually significant regions is challenging since it is in contradiction with the imperceptibility of the watermark.

Laplacian Removal Attack Operator is another signal processing technique to remove the watermark, which is proposed by Barnett [1]. The attacked image is $M_{attacked} = M' - \alpha(L_p - L_n)$, where L_n is the resulting image of M' convoluted with a Laplacian 3 * 3 mask, and convolute L_n again with the Laplacian mask to get L_p. Experiments [1] show that the attack in frequency domain is more effective and it is called Frequency mode Laplacian removal attack.

C) Watermark Estimation Through Detector Analysis

It is commonly known that the existence of watermark detector on the client side is risky. It is because reverse engineering can be done so that the watermark can be removed perfectly. This may be in fact very difficult depending on the watermark and software design. However, with the existence of the detector in the client side, a more simple but time consuming attack can remove the watermark without any knowledge about the watermarking techniques [3][6]. First, the attacker modifies the watermarked image until the detector just responds that there is no watermark embedded, no matter how the resultant image is distorted. One possible modification is to reduce the contrast of the image, until the detector is not able to detect the watermark. The attacker then increases or decreases the luminance pixel by pixel, until the watermark appear to the detector again. This gives the attacker the knowledge about how sensitive the detector responds to each individual pixel and the attacker can then estimate a combination of pixel values that has the largest influence on the detector for the least distortion to the image.

D) Watermark Copy Attacks

The Copy Attacks is proposed by M. Kutter et. al. [8]. This watermark copy attack, falling in the category of Ambiguity Attacks, aims not to remove or destroy the watermark, but to "copy" the watermark from one watermarked image to another un-watermarked image. This attack is based on estimating the embedded watermark of the real watermarked image, and adaptively adding to the attacker's image in order to keep the imperceptibility. This gives a new challenge for those applications using watermark to prove genuineness and control access. It used our discussed watermarking model: $M' = M + f(M,W)$. Considering the watermark is the noise, we can then estimate the watermark U_E as $U_E = M' - M_E$, where M_E is the estimate of the original. There are two proposed model to estimate the watermark: Maximum likelihood (ML) estimation and Maximum a posteriori Probability (MAP) estimation. MAP works only if we have the prior statistic of the un-watermarked original image, which is not a usual case. For the ML estimation, the predicted original is simply the local mean of the watermarked image if the watermark is Gaussian distributed.

E) Attacks on Copy Control Applications

Some copy control applications use watermark to mark the media as "copy once" or "no copy allowed", etc. For example, a watermark in a video can tell a compliant recorder not to record the video. This compliant recorder (watermark detection enabled) will record a copyrighted video, however, most likely it should also be able to record non-copyrighted materials. In such an application domain, attackers can then scramble the copyrighted video before recording it to the recorder, such that the recorder is not able to detect the watermark information and regard it as a non-copyrighted video [3]. This kind of detection disabling attacks is especially suitable to handle media in digital format, since perfect scrambling and de-scrambling can be done.

F) SWICO Attacks

Single Watermarked Image Counterfeit Original (SWICO) Attacks [4], or sometime known as IBM Attacks, is another Ambiguity Attacks proposed by Craver et. al. This is again a very simple attack, which challenge the

rightful ownership problem of many watermarking schemes. The idea of this attack is to fabricate a Counterfeit Original from the watermarked real image. This can be simply achieved by *removing* a randomly selected watermark *U'*, which is the attacker's watermark, from the watermarked real image *M* forming a fake original *M'*, instead of embedding one. As a result, the real watermarked M_w contains the fake original *M'* and there is no method to find out who embed the watermark first. As we discussed before, this kind of attacks exists because there is no definition of watermark and watermarking technique.

6. PIXEL REALLOCATION ATTACK

In this section, we propose a new attack, Pixel reallocation attack, and experimental results are presented in the next section. The proposed attack is a hybrid approach, which use two different techniques to remove the watermark from the watermarked image. This hybrid approach is proved to lead to a better resultant image in term of the visual quality. Single technique approach needs a larger distortion to the image in order to attack the watermarked image successfully.

The proposed attack demonstrates that new attack can be found readily, in contrast to the development of counter solutions. New attack can be a combination of ordinary attacks and applied together to the watermarked media. By such, watermark designer is hard to implement "anti-attack" techniques to the watermark design without the prior knowledge of this class of hybrid or cocktail attacks.

Pixel reallocation attack

Pixel reallocation attack is a simple attack that aimed to decorrelate the embedded watermark with the original watermark. Since many watermarking detections correlate the testing image with the target watermark, it will not work once we decorrelate the embedded watermark. For example, the geometrical transformation attacks, which we discussed in Section 4, desynchronize the correlation detector with the testing image that leads to detection failure. However, by inserting a template or grid into the watermarked image can make inverse transformation possible and the detection can still succeed. If we can apply the transformation to every single pixel locally, independently and randomly, inverse transformation will not be possible and the attack will be successful, since the attacked image is not correlated to the original watermark.

The proposed attack is as follow:

1) Randomly select a neighbor pixel within a window size $\pm W$
2) Replace the current pixel with the randomly selected neighbor pixel if their absolute difference is less then threshold *T*
3) If their absolute difference is larger then threshold *T*, repeat Step 2
4) Repeat Step 1 and 2 for all pixels
5) Filter the attacked image with a standard 3*3 median filter

The need of the threshold *T* is to limit the distortion results from the attack and to ensure that a pixel will not be replaced with a pixel, which is on the other side of an edge, since pixel values on different side of an edge have a big difference in general. Median filter can help to remove the impulsive noise due to the reallocation and remove the embedded watermark further.

7. RESULTS

In the experiment, Lena and Peppers (both are 512X512, 8bit gray scale), are watermarked with the well-known digital image watermarking software, Digimarc, with a "watermark durability" set to medium (value 8). The proposed attack is then applied to these watermarked images and processed with the watermark detection. Figure 1 and 2 shows the original and watermarked images.

Different parameters of the attack are used. The results show that, as expected, the success of an attack depends greatly on the threshold the windows size. Since our attack is designed to keep sharp edges of image, resultant images show that the proposed attack works well in those texture or high frequency regions, like the fur of Lena's hat, besides flat regions.

The most significant result draws from the experiment is that by applying two watermark attacks together (reallocation and median filtering) would introduce less distortion on the image for a successful attack.

For the image Lena, if we apply pixel reallocation with median filter, the attack removes the watermark, with the PSNR of the resultant image is equal to 32.16 dB. If we apply pixel reallocation attack without median filter, we need to increase the threshold to 20 in order to remove the watermark from the image, such that the PSNR drops to 29.40 dB. For the image Peppers, as the same as Lena, a larger distortion induced on the attacked image if we use one attack technique only.

Figure 1: Original Lena (left) and Watermarked Lena with Digimarc (right)

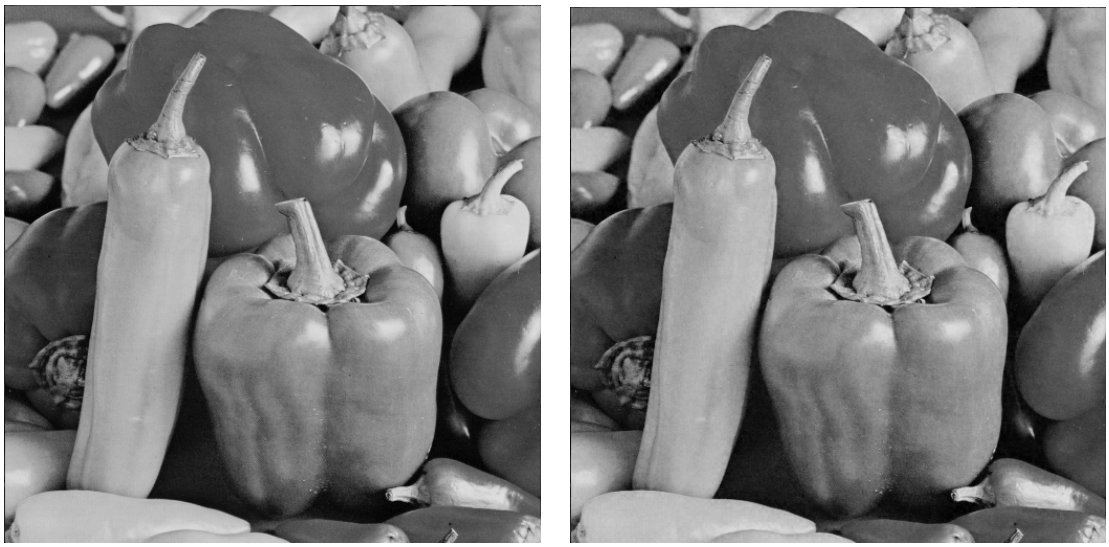

Figure 2: Original Peppers (left) and Watermarked Peppers with Digimarc (right)

Figure 3: Successful attacked images of Lena.
Left: pixel reallocation and median filtered, window size 8, threshold 15, PSNR 32.16.
Right: pixel reallocation only, window size 8, threshold 20, PSNR 29.40

Figure 4: Successful attacked images of Peppers.
Left: pixel reallocation and median filtered, window size 8, threshold 15, PSNR 32.27.
Right: pixel reallocation only, window size 8, threshold 20, PSNR 29.22

Lena, 512*512, 8bit gray scale					
Window Size (±W)	Threshold	With median filter		Without median filter	
		PSNR	WM detected	PSNR	WM detected
N/A (Original)	N/A	N/A	×	N/A	×
N/A (Watermarked)	N/A	39.55	✓	39.55	✓
8	5	34.09	✓	37.20	✓
8	10	33.21	✓	33.56	✓
8	15	32.16	×	31.11	✓
8	20	31.21	×	29.40	×
8	25	30.35	×	28.09	×
8	30	29.61	×	27.04	×
16	5	34.05	✓	37.06	✓
16	10	33.02	✓	33.21	✓
16	15	31.80	✓	30.56	✓
16	20	30.63	×	28.67	✓
16	25	29.56	×	27.28	×
16	30	28.74	×	26.14	×

Table 1: Results for attack of the image Lena

Peppers, 512*512, 8bit gray scale					
Window Size (±W)	Threshold	With median filter		Without median filter	
		PSNR	WM detected	PSNR	WM detected
N/A (Original)	N/A	N/A	×	N/A	×
N/A (Watermarked)	N/A	39.85	✓	39.85	✓
8	5	34.00	✓	37.40	✓
8	10	33.25	✓	33.48	✓
8	15	32.27	×	30.94	✓
8	20	31.38	×	29.22	×
8	25	30.57	×	28.02	×
8	30	29.91	×	27.02	×
16	5	33.98	✓	37.19	✓
16	10	33.06	✓	33.15	✓
16	15	31.85	✓	30.40	✓
16	20	30.68	×	28.50	✓
16	25	29.68	×	27.07	×
16	30	28.81	×	25.96	×

Table 2: Results for attack of the image Peppers

8. CONCULSIONS

In this paper, we characterized and discussed different types of watermark attacks. We showed that the development of a robust watermarking technique is challenging and difficult. There are no existing watermarking techniques, which are robust to all different types of attacks that we discussed here. Although more and more new watermarking techniques are being proposed, there are also many new watermark attacks emerging. Like the Pixel reallocation attacks that we proposed in this paper, commercial watermarking software cannot survive under this simple attack. We can foresee that such a new attack can be found easily by combining two or more basic attacks or by selecting some simple manipulations to the media arbitrarily with minimum effort; on the other hand, counter solutions to these new attacks are difficult to developed, since we do not have prior knowledge about these attacks. To claim to be robust, watermarking techniques must overcome all these attacks. However, this will also increase the complexity of the techniques and may add more distortion to the watermarked media. The cost of such watermarking process may then be, in some senses, higher then the value of the embedded information, such as the copyrights. We may need to redefine the scope, application and usage of watermarking. Other watermarking application, like fragile watermarking, watermark for authentication, data embedding, etc., might be more feasible and have a better valuation then robust watermarking, by the fact that attacks or post processing to these applications are more predictable and better defined.

REFERENCES

[1] R. Barnett, and D. Pearson, "Attack Operators for Digitally Watermarked Images", *Proc. of IEE Visual Image Signal Process*ing, Vol. 145, No. 4, pp. 271-279, Aug 1998.

[2] I.J. Cox, J. Killian, F.T. Leighton, T. Shamoon, “Secure Spread Spectrum Watermarking for Multimedia”, *IEEE Trans. on Image Processing*, Vol. 6, No. 12, pp.1673-87, Dec 1997.

[3] I.J. Cox, M.G. Linnartz, "Some General Methods for Tampering with Watermarks" *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 4, pp. 587-593, May 1998.

[4] S. Craver, N. Memon, B.L. Yeo, "Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks, and Implications", *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 4, pp. 573-586, May 1998.

[5] F. Hartung, J.K. Su, B. Girod, "Spread Spectrum Watermarking: Malicious Attacks and Counterattacks", *Proc. of SPIE Conf. on Security and Watermarking of Multimedia Contents*, Vol. 3657, pp. 147-158, Jan 1999.

[6] T. Kalker, J.P. Linnartz M.V. Dijk, "Watermark Estimation Through Detector Analysis" *Proc. of IEEE Int. Conf. on Image Processing.* Vol.1, pp.425-9, Oct 1998.

[7] M. Kutter, "Watermark Resisting to Translation, Rotation and Scaling", *Proc. of SPIE Multimedia System and Applications*, Vol. 3528, pp. 423-431, Nov 98.

[8] M. Kutter, S. Voloshynovskiy, A. Herrigel, "The Watermark Copy Attack", *Proc. of SPIE: Security and Watermarking of Multimedia Content II*, Vol. 3971, Jan 2000.

[9] S.H. Kwok, K.F. Tsang, "Adaptive Audio Watermarking Scheme for Copyrights Protection", *Proc. of Asia Pacific Conf. on Information Systems 2000*, pp225-37, June 2000.

[10] J.J.K. O'Ruanaidh, T. Pun, "Rotation, Scale and Translation Invariant Digital Image Watermarking", *Proc. of IEEE Int. Conf. on Image Processing*, Vol. 1, pp.536-9, 1997.

[11] M.D. Swanson, B. Zhu, A.H. Tewfik, L. Boney, “Robust Audio Watermarking Using Perceptual Masking”, *Signal Processing*, Vol. 66, No. 3, pp.337-55, May 1998.