Image Watermarks and Counterfeit Attacks: Some Problems and Solutions

Mahalingam Ramkumar, Ali N Akansu
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
New Jersey Center for Multimedia Research
University Heights, Newark, NJ 07102.

Abstract

Several watermarking schemes have been proposed in recent years with proven robustness to many types of intentional and unintentional signal processing attacks. However very few investigators [1], [2] have addressed the question of "unambiguous resolution of ownership with digital watermarks". In this paper we extend the work of Craver *et. al.* [1]. We show how a pirate can engineer an attack against the most robust scheme proposed in [1] with reasonable computational complexity. We then propose a improvement on that scheme which increases the complexity for engineering a successful attack by a factor of 10^{100} to 10^{200}, thus making it virtually attack proof. We also indicate other restrictions to be placed on watermarking schemes to make the watermark more robust to intentional attacks.

I. Introduction

DIGITAL watermarking is a means of protecting multimedia data from intellectual piracy. It is achieved by imperceptibly modifying the original data to insert a "signature". The signature is extracted when necessary to show proof of ownership. In this paper, we restrict ourselves to watermarking digital images.

Let I be the original (cover) image. A watermark embedding function $\mathcal{E}$ inserts a watermark S in the image I to obtain the watermarked image $\hat{I}$.

$$\hat{I} = \mathcal{E}(I, S) \tag{1}$$

The existence of the watermark S in an image $\tilde{I}$ is checked by a detector function $\mathcal{D}$. Watermark detectors can be broadly classified into two categories. Some detectors need the original image I to check for the presence of the signature S in $\tilde{I}$. Such schemes, are called *cover image escrow* schemes. On the other hand schemes that do not require the original image for detection of the signature are called *oblivious detection* schemes. We shall term the output of the detector function,

$$s_d = \begin{cases} \mathcal{D}(\tilde{I}, S, I) & \text{cover image escrow schemes} \\ \mathcal{D}(\tilde{I}, S) & \text{oblivious detection schemes} \end{cases} \tag{2}$$

as the *detection statistic*. The detection statistic is an indication of the *degree of certainty* with which the signature S is detected in the image $\tilde{I}$.

Typically, the embedding function adds a random sequence S to I in a transform domain. The detection statistic s_d may be the projection of the signature S onto the transform coefficients of the image. Though other types of embedding and detecting functions are possible, in all examples in this this paper, for simplicity, we assume that the signature is detected by correlative processing (projection of the signature onto the transform coefficients of the image). However, the improvements we propose to watermarking schemes are applicable for all types of existing watermarking schemes.

From the early fragile watermarking schemes [3] that modified only the LSB's of digital multimedia signals, state-of-the-art image watermarking schemes [2], [4], [5], [6] have come a long way in terms of robustness to intentional and unintentional signal processing attacks. Though most of the methods fail to address the question of whether their schemes would hold in a court of law, it does not imply that those schemes would be of no use. They could be modified slightly, to make them acceptable (in the lines of modifications to Cox's scheme [4] by Craver *et. al.* in Ref. [1]). With a wealth of excellent schemes (for both cover image escrow and oblivious detection) available for watermarking, it is probably time to revisit the question of the ability of a watermarking scheme to unambiguously resolve rightful ownership.

We begin by reviewing schemes suggested by Craver *et. al.* in [1]. We then show how the most robust (in terms of resistance to counterfeit attacks) scheme suggested in [1] may still be inadequate. We finally propose some modifications to that scheme to make it more robust.

II. Problem Statement

It has almost become a tradition in watermarking literature, to have Alice as the originator / creator of the image and Bob as the aspiring forger / pirate. Alice is the creator of the original image I. She adds two signatures S_A and S_n in the image I to create the watermarked image. The signature S_A would identify the owner (Alice). The signature S_A is added to all copies of the image. S_n would probably be a signature that would identify the "serial number" of a copy. Therefore, S_n would be different for different copies. Usually the signature is added in some transform domain (like DCT, DFT, Hadamard, or wavelet transform). Let I_t denote the transform domain coefficients of I [1]. The watermarked image $\hat{I}$ is obtained as

$$\begin{aligned} \hat{I}_t &= I_t + S_A + S_n \\ \hat{I} &= \mathcal{T}^{-1}(\hat{I}_t), \end{aligned} \tag{3}$$

where $\mathcal{T}$ is the transform employed. Though Eq. (3) may not be a very general description of possible watermarking schemes, it encompasses most of the state-of-the-art schemes.

Now the following sequence of events occur:

- Alice sells a copy of $\hat{I}$ to Bob.
- Bob redistributes *illegal* copies of $\hat{I}$. He may have modified the illegal copies to a certain extent. Let the copies made by Bob be $\tilde{I}_1 \cdots \tilde{I}_k$.
- Alice 'stumbles' upon an illegal copy (say $\tilde{I}_l$). She extracts the signature S_n from the illegal copy and finds the serial number of the copy sold to Bob.
- Alice decides to sue Bob for breach of contract.

Now Alice has to prove in a court of law that

- The image in question, viz. $\tilde{I}_l$, is owned by her.
- Bob is responsible for redistributing the illegal copy. Or the illegal copy originated from the image $\hat{I}$ sold to Bob.

Now that the background has been established we shall see how Bob can invalidate Alice's claims, and what Alice should do to make her claims acceptable in court.

III. Different Watermarking Schemes and Counterfeit Attacks

In this section we shall see how different watermarking schemes resolve (or fail to resolve) true identity of the owner.

A. *Scheme I*

In this scheme

- Alice chooses an arbitrary signature. But the signature is registered with some *relevant authority* (lets call them the Watermarking Authority), and approved.
- The signature is detected as follows:

[1] I_t may be a subset of the transform domain coefficients of I

– The transform coefficients of Alice's original image I and the image of questionable origin $\tilde{I}_l$ are obtained

$$\begin{aligned} I_t &= \mathcal{T}(I) \\ \tilde{I}_{l_t} &= \mathcal{T}(\tilde{I}_l) \end{aligned} \tag{4}$$

– The difference $\tilde{I}_{l_t} - I_t$ is normalized, and correlated with the sequence S_A. If the result of the correlation is above a threshold, then it is assumed that the signature is present. (If the transform $\mathcal{T}$ is linear, then $\tilde{I}_{l_t} - I_t$ can be obtained as $\mathcal{T}(\tilde{I}_l - I)$).

While Alice extracts the signature S_A from $\tilde{I}_l$, Bob seems undaunted. Bob can very easily counterfeit Alice's claim. His argument is as follows:

- $\hat{I}$ is the original image. Bob was not interested in watermarking his original image.
- Alice "stole" a copy of $\hat{I}$.
- Alice registered her signature S_A. Then she subtracted S_A from her stolen copy of $\hat{I}$ to create her "original" image I.

Bob, however, cannot *prove* that he is the owner. But by establishing reasonable doubt, he escapes conviction. Now that the even the owner of the image is in question, the court is not interested in checking the serial number (the signature S_n).

B. Scheme II

The problem with the first scheme is that the original image is used in the signature extraction problem (or so Alice thinks, incorrectly). Therefore, she modifies her scheme as follows:

- She registers her signature as in Scheme I
- The signature is extracted as follows
  – The transform coefficients of $\tilde{I}_l$ are obtained.
  – The sequence S_A is correlated with $\tilde{I}_{l_t}$ (the signature is detected without subtracting the original image).

But this is still not good enough to implicate Bob. His claim is as follows:

- $\hat{I}$ is the original image.
- Alice stole $\hat{I}$. She then engineered a signature that yielded a high correlation with $\hat{I}_t$, and registered it as S_A.
- As any modified version of $\hat{I}$ is still likely to be very similar to $\hat{I}$, S_A would yield a high correlation with all images derived from $\hat{I}$.
- If Alice demonstrates the lack of correlation of S_A with I (I_t to be more specific), Bob can still claim that Alice subtracted S_A from $\hat{I}_t$ to obtain I_t (and hence I).

Thus Bob succeeds in countering Alice's claim.

C. Scheme III

Alice takes suggestions from Craver *et. al.* [1]. She realizes the problem is because her signature is not *constrained*. So the Watermarking Authority places the following restrictions:

- A *fixed* hash function $\mathcal{H}$ is to be used. The function $\mathcal{H}$ operates on the original image I to produce a 'seed'. This seed is used by a (also fixed) random sequence generator to generate the signature sequence (say Gaussian random sequence). No restriction is placed on the length of the signature sequence to be used.
- Any decomposition can be used for embedding. But strict guidelines will be placed on how the coefficients may be reordered. Arbitrary ordering of the transform coefficients will not be allowed.

Note that the restrictions placed are very practical. It just involves fixing the hash function and the random sequence generator. Moreover, it does not involve an on-going involvement of the Watermarking Authority with the watermark extraction process.

Fig. 1. Left : Original Goldhill image. Center: Goldhill image obtained by modifying the histogram. Though both images look similar, and are of good visual quality, the difference between the two images in terms of PSNR is 20 dB. Right: Image obtained after StirMark.The difference between the two images in terms of PSNR is 19 dB.

With these restrictions Alice embeds her watermark in I to get $\hat{I}$. For detection she may subtract the original I_t from $\tilde{I}_{l_t}$ before performing correlation with S_A. S_A is obtained from the fixed hash function $\mathcal{H}$ as

$$S_A = \mathcal{H}(I). \tag{5}$$

Now Bob cannot claim that Alice stole $\hat{I}$, engineered a signature and subtracted it from $\hat{I}$ as in Scheme I. This is due to the fact that

$$S_A = \mathcal{H}(I) \neq \mathcal{H}(\hat{I}) \tag{6}$$

However a counter claim for this scheme may not be as difficult as it appears at first sight. In the next section we show how Bob, with some effort, can create ambiguity in the proof of ownership.

IV. Countering Scheme III

Bob does the following

- Change $\hat{I}$ significantly, in the mean-square-error sense while maintaining the "visual similarity" between the original $\hat{I}$ and the resulting (modified) image $\hat{I}_m$.
- Employ some standard watermark attack software on $\hat{I}_m$ so that Alice cannot detect her signature with a very high degree of certainty.

Though there are many ways to do accomplish step 1, the simplest is probably to obtain $\hat{I}_m$ by modifying the histogram of $\hat{I}$. As an example the original Goldhill image (256×256 pixels) is shown in Figure 1 - Left. Figure 1 - Center shows the modified Goldhill image obtained by reshaping the histogram. Though both images are very similar and are of good visual quality, the difference in terms of PSNR between the two images is 20 dB! Actually using a watermark attack software like Stirmark [2](which apart from other things, simulates the conditions that an image undergoes during printing-photocopying-rescanning cycle) accomplishes both steps at the same time. Application of Stirmark on 15 256×256 test images yielded resulting images, of reasonably good perceptive quality, though the difference in PSNR between the original and the modified image was around 19 to 20 dB on an average (Figure 1 - Right). Let I_d be the difference image

[2]available for download from `http://www.cl.cam.ac.uk`

$$I_d = \hat{I}_m - \hat{I} \tag{7}$$

The total power of I_d much larger than that of the signature S_A added by Alice. In other words,

$$\sum_{i=1}^{M1}\sum_{j=1}^{M2}(I(i,j) - \hat{I}_m(i,j))^2 >> \sum_{i=1}^{M1}\sum_{j=1}^{M2}(I(i,j) - \hat{I}(i,j))^2, \tag{8}$$

where $M1$ and $M2$ are the image dimensions, and

$$\sum_{i=1}^{M1}\sum_{j=1}^{M2}(I(i,j) - \hat{I}(i,j))^2 = \sum_{k=1}^{N} S_A(k)^2. \tag{9}$$

From Eqs. (7), (8) and (9)

$$I_d = \hat{I}_m - \hat{I} \approx \hat{I}_m - I \tag{10}$$

Now Bob derives his "original" image from $\hat{I}_m$. Before we see how he does that, note that the hash function $\mathcal{H}$ maps different images to (possibly) different seeds. For example if all the images in the world were of size 256×256 and restricted to 8 bits per pixel, there are still $2^{256\times256\times8}$ possible images. Though $\mathcal{H}$ would map the space of images to a (comparatively) very restricted 'space' of seeds, the space of seeds should still be large enough so that the probability that different signatures are correlated is very small. Two 'obviously' different images having the same signature is not likely to create a problem. The problem only arises when images are 'similar'. So it is important that the (fixed) hash function generates different seeds especially when the images are 'similar' So the hash function would be required to "respond" to the LSBs of image more than to the MSBs. This works to Bob's advantage.

Bob could probably generate enough (different) signature sequences from the image $\hat{I}_m$ just by tweaking 1-2 LSBs of the image pixels. But when he does that the resulting image is still very close to $\hat{I}_m$. So he correlates every signature sequence with the fixed I_{d_t}, which is the transform domain equivalent of I_d (if Bob has enough computing resources, he may even obtain the new I_{d_t} every time $\hat{I}_1$ is modified). Whenever a particular tweaking of the bits results in a signature sequence with satisfactory correlation with I_{d_t}, he stops. He calls the resultant image $I_m \approx \hat{I}_m$ as his "original" image. If S_B is the signature generated from I_m, and S_B has a reasonable correlation with (the transform domain equivalent of) $\hat{I} - I_m$, then it can also be expected to have high correlation with (the transform domain equivalent of) $I - I_m$. So Bob can demonstrate the presence of his signature in I! Note that making $I_m - \hat{I}$ large swamps out the difference between I and $\hat{I}$.

A. Computational Complexity of the Attack

Let S_b be the random Gaussian sequence generated by Bob (using the hash function on his "original image" I_m). The detection statistic is obtained as

$$s_d = \langle S_b, I_{d_t} \rangle \tag{11}$$

As S_b is a Gaussian sequence, and I_{d_t} is fixed, s_d is a linear combination of many Gaussian variables, and hence, Gaussian. Let σ_d be the standard deviation of s_d.

Now the question is, what should be the value of the detection statistic s_d to demonstrate the presence of the watermark in the image in question? Alternately, what is the *measure of certainty* with which the watermark is detected, given the value of s_d? One way to quantify the measure of certainty is through an estimate of the standard deviation of s_d. For example, if we desire a probability of error, P_e in detection of the signature to be less than 10^{-9}, we choose the detection threshold as

$s_d = 6\sigma_d$. As $s_d \sim \mathcal{N}[0, \sigma_d^2]$, the probability that $s_d > 6\sigma_d = Q(6) \approx 10^{-9}$. Alternately, if the detection threshold $s_d = 6\sigma_d$, then one can expect one out of $\frac{1}{10^{-9}}$ randomly generated signatures to yield a detection threshold equal to or greater than $6\sigma_d$.

State-of-the art watermarking schemes [5], [2] are capable of detecting the signature with very high degrees of certainty ($P_e < 10^{-30}$) even when the image has undergone very low quality JPEG (say 10 % quality). But after carefully planned attacks on the watermark, the degree of certainty may reduce to the order of $10^{-6} - 10^{-8}$. For example the scheme in [5] detects the watermark in a image after low quality printing-photocopying-rescanning cycle with $P_e < 10^{-6}$. After application of StirMark however, the detection statistic yields $P_e \approx 10^{-2}$. Alice, therefore may not be able to detect her signature in I_m with a high degree of certainty. It should be appreciated here that as long the detection statistic of Bob's signature in I is comparable to the detection statistic of Alice's signature in I_m, it does not help Alice in any way to obtain much higher detection statistic in $\tilde{I}_l$ than Bob can.

Lets assume that Alice, using a very sophisticated watermarking scheme manages to detect her signature in I_m with $P_e < 10^{-9}$, (or $s_d > 6\sigma_d$). To obtain a comparable detection statistic of his signature in I, Bob has to *search* 10^9 sequences on an average before obtaining a suitable signature. This is certainly computationally feasible.

V. A Modification to Scheme III

In this section we propose a modification to Scheme III which increases the computational complexity of the attack by a factor of over 10^{100}. The only differences between the scheme III and the proposed scheme are

- The watermark should be detected *without subtracting* the original image. But the original image is still necessary because the seed is obtained from the original image.
- The signature should yield a high correlation (detection statistic) with with the image in which the signature is to be detected. In addition, the signature should yield a *low* correlation (detection statistic) with the original image.

Let s_i be the detection statistic obtained by correlating the transform coefficients I_t of the actual original image I (in which Bob proposes to show his signature). Let σ_i be the standard deviation of s_i. To show his signature in the image I with the same degree of certainty as in Scheme III ($P_e < 10^{-9}$), the signature should be chosen such that $s_i > 6\sigma_i$. In addition, the same signature should also yield a *low correlation* with Bob's "original" image I_m. Let s_{i_m} be the statistic obtained by correlating the signature with the transform coefficients of I_m. Obviously the detection statistics s_i and s_{i_m} are not independent. As I and I_m are still more "similar" than "not similar", one would expect a random sequence that yields a high correlation with I to also yield a high correlation with I_m. This makes it extremely difficult for Bob to engineer a signature. Now,

$$\mathrm{Prob}[(s_i > 6\sigma_i) \cap (-\delta < s_{i_m} < \delta)] < \mathrm{Prob}[(s_i - s_{i_m}) > 6\sigma_i - \delta] \tag{12}$$

Alternately,

$$\text{Probability of Bob succeeding} < \mathrm{Prob}[(s_i - s_{i_m}) > 6\sigma_i - \delta] \tag{13}$$

For the example image the statistics of $s_o = s_i - s_{i_m}$ (which is also Gaussian) was found to have standard deviation σ_o of about $0.25\sigma_i$. The probability of finding a signature to engineer the counter claim drops from $Q(6) \approx 10^{-9}$ for Scheme III to $Q(6\frac{\sigma_i}{\sigma_{i_m}}) \approx 10^{-127}$ for the suggested scheme. In other words, Bob would have to search an average of 10^{127} sequences before finding a suitable signature! To engineer a signature that will be detected with $P_e < 10^{-5}, 10^{-8}, 10^{-12}, 10^{-20}$, Bob has to search an average of $10^{65}, 10^{111}, 10^{174}, 10^{300}$ signatures, respectively!

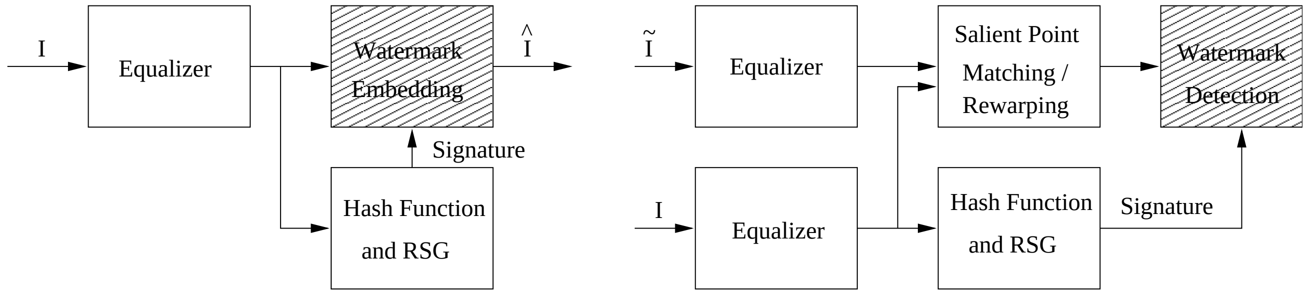

Fig. 2. Watermark Embedding and Detection Protocol

VI. Overcoming Attacks on Watermarks

With the advent of watermark attack software like StirMark, it is becoming increasingly clear that watermarking schemes based on linear addition and detection of signature are bound to fail. The linear addition methods depend on the assumption that if the image is altered significantly in the mean-square sense, then the quality of the resultant image would be so poor that it would not warrant a ownership claim. However, modifying the histogram (Figure 1) and warping images imperceptibly (like StirMark does in Figure 2) can significantly alter the image in the mean-square sense while maintaining visual similarity. One could still use linear watermarking methods effectively against attacks that modify the histogram by permitting reshaping the histogram of the image in question to match the histogram of the original image before detecting the signature. Similarly a good algorithm for detecting *salient points* of the original image and those of the image in question may be used to re-warp the image so that the salient points match, before the signature is detected.

Of course, the algorithm for reshaping the histogram and identifying the salient points to re-warp the image, should be fixed, and global (permitting modifications of these schemes would provide the pirate with additional degrees of freedom to engineer signatures).

VII. Suggestions for Regulatory Agencies

It is very important that a standards committee be established at the earliest for regulating and controlling watermarking protocols. We suggest the following list of restrictions to be placed on watermarking schemes, in order to make them resolve rightful ownership unambiguously. The overall protocol for watermark embedding and detection are shown in Figure 3. All the unshaded blocks in Figure 3 are fixed. Only the watermark embedding function and detecting function will depend on the watermarking scheme.

- A fixed algorithm for equalizing histogram. The signature is added to the original image after equalizing its histogram. The histogram of the image in question is equalized (using the same equalizer) before performing detection of the signature.
- A fixed algorithm for determining significant points and re-warping the image if necessary.
- Fixed hash function $\mathcal{H}$ to be used. The hash function could be made computationally intensive to further discourage engineering of digital signatures. The hash function operates on the (histogram equalized) original image I to produce the seed $\mathcal{H}_I$.
- The seed $\mathcal{H}_I$ is input to a *fixed random sequence generator* $\mathcal{G}$ to generate the signature sequence S_I.

$$S_N^d = \mathcal{G}(\mathcal{H}_I, N, d) \tag{14}$$

is the complete set of sequences that could be generated by $\mathcal{G}$. For a fixed I, the only parameters that can be changed are N - the length of the sequence, and d - the probability distribution. Probably d could take two options - Gaussian and Uniform. Another useful option for d might be to generate a list of integers from $1 \cdots N$ in a random order. This may be used for reordering the image coefficients if the algorithm calls for it. No restriction is placed on the length N.

• Any decomposition of the original image can be used. If decompositions are generated from random sequences only one from the set of possible sequences S_N^d can be used. If the watermarking algorithm calls for a random sequence at any stage of the watermark embedding / extraction process, only random sequences S_N^d are permitted.
• Signature to be extracted from the image without subtracting the original image.
• High correlation (detection statistic) of the signature with the image in which the existence of the signature is checked, *and* low correlation between the signature and the original image. (This restriction may result in a highly improbable scenario of the signature generated by hashing the original image having a high correlation with the image. Under this scenario, the originator will be forced to tweak a few bits of his original image to generate a different signature)

It should be mentioned here that the word 'correlation' is used rather loosely here. The proposal does not limit itself only to schemes in which the signature is detected by correlative processing. For example, in [5] some low frequency DCT coefficients are scrambled by a random cyclic all-pass filter. The detection statistic is obtained by counting the difference between positive and negative coefficients. The only restriction the proposal places on the scheme above is how the seed is obtained and the corresponding random sequence to be used to generate the all-pass filter coefficients. For schemes that do not use correlative processing, substitute the more general 'detection statistic' instead of 'correlation' in the list above. To our knowledge any existing oblivious detection watermarking scheme can be modified to meet the requirements of the proposed scheme.

VIII. Conclusions

In this paper, we have introduced a modification to the watermarking scheme proposed by Craver *et. al.* to significantly increase the robustness of watermarks to counterfeit attacks. We offer a complete list of restrictions to be placed on watermarking schemes so that the final scheme meets the end requirement, viz. unambiguous resolution of ownership. The restrictions, to our knowledge, does not limit the applicability of any existing oblivious detection watermarking scheme.

References

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

[2] W.Zeng, B. Liu, "On Resolving Rightful Ownerships of Digital Images by Invisible Watermarks", IEEE Conference on Image Processing, vol **1** pp 552-555, Santa Barbara, CA, Oct 26-29, 1997.

[3] W. Bender, D. Gruhl, N. Mormato, "Techniques for Data Hiding", SPIE **2420**. Feb. 1995.

[4] I.J. Cox, J. Kilian, F.T. Leighton, and T.G. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia", IEEE Transactions on Image Processing, **6** (12) pp 1673-1687, 1997.

[5] M.Ramkumar, A.N. Akansu, "A Robust Scheme for Oblivious Detection of Watermarks / Data Hiding in Still Images", SPIE's Symposium on Voice, Video and Data Communication (VV-06), Boston, MA, 2-5 Nov. 1998.

[6] M.D. Swanson , B. Zhu , and A.H. Tewfik , "Transparent Robust Image Watermarking," Proc. of the 1996 IEEE Int. Conf. on Image Processing , Vol. III, pp 211-214, 1996

[7] N.Memon, P.W. Wong, "A Buyer-Seller Watermarking Protocol", IEEE Workshop on Multimedia Signal Processing (MMSP-98), Dec 7-9, Los Angeles, California, USA, pp 278-283, 1998.