CERIAS Tech Report 2007-44

A WATERMARKING TECHNIQUE FOR DIGITAL IMAGERY: FURTHER STUDIES

by Raymond B. Wolfgang and Edward J. Delp

Center for Education and Research in
Information Assurance and Security,
Purdue University, West Lafayette, IN 47907-2086

A WATERMARKING TECHNIQUE FOR DIGITAL IMAGERY: FURTHER STUDIES

Raymond B. Wolfgang and Edward J. Delp

Video and Image Processing Laboratory (*VIPER*)
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, 47907-1285
USA

Abstract

The growth of networked multimedia systems has created a need for the copyright protection of digital images. Copyright protection involves the authentication of image ownership and the identification of illegal copies of a (possibly forged) image. One approach is to mark an image by adding an invisible structure known as a digital watermark. In this paper we further study techniques for marking images introduced in [1]. In particular, we describe how our techniques withstand random errors. We also provide more details relative to our verification procedure. Finally, we discuss the recently proposed IBM attack.

Keywords: Digital Watermarking, Multimedia Security, Copyright Protection

1. Introduction

The recent growth of networked multimedia systems has increased the need for the protection of digital media. This is particularly important for the protection and enforcement of intellectual property rights. Digital media includes text, digital audio, images, video and software. Many approaches are available for protecting digital data; these include encryption, authentication and time stamping. We propose to study techniques for image authentication and forgery prevention known as watermarks.

This work was partially supported by a grant from the AT&T foundation. Address all correspondence to E.J. Delp, ace@ecn.purdue.edu, http://www.ece.purdue.edu/~ace, or +1 765 494 1740.

Techniques are needed to prevent the copying, forgery and unauthorized distribution of images and video. Without such methods, placing images on a public network puts them at risk of theft and undetected alteration. The basic scenario is as follows: A user has created an electronic image at some effort and expense, and wants to make it available on a communications network. When unauthorized copies or forgeries of the image appear elsewhere on the network, the user needs to prove that the image belongs to them. One also needs to determine if and by how much the image has been changed from the original. Image protection algorithms must provide:

1. Copy detection to identify unauthorized copies of an image.
2. Content authentication to verify the content of a copy of an image, since the copy may have been forged or filtered.
3. Owner authentication to prove that the user is the true owner of the image.

Other requirements could be:

1. Chain-of-custody determination
2. Time stamping to verify when an image was created and/or viewed.

2. Background

A watermark is a secret code or image incorporated into an original image. The use of perceptually invisible watermarks is one form of image authentication. A watermarking algorithm consists of three parts: the watermark, the marking algorithm and the verification algorithm. Each owner has a unique watermark. The marking

algorithm incorporates the watermark into the image. The verification algorithm authenticates the image, determining both the owner and the integrity of the image.

Many digital watermarking techniques rely on random sequences incorporated into to an image's spatial or spectral representation. One approach adds a modified maximal-length linear shift register sequence (m-sequence) to the pixel data. They identify the watermark using correlation techniques [2,3]. Watermarks can also modify the image's spectral or transform coefficients directly. These algorithms most often modulate DCT coefficients according to a sequence known only to the owner [4]. A different spectrum-based technique passes the image through a sub-band filter before marking an image [5]. Many of these watermarking techniques depend on the image content; the techniques increase the level of the watermark in the image while maintaining the imperceptibility of the mark [6,7]. Other watermarks also use the Human Visual System [8]. Visible watermarks also exist; IBM has developed a proprietary visible watermark to protect images that are part of the digital Vatican library project [9]. Most of these techniques may be used in combination with each other.

M-sequence spatial methods tolerate errors to an image, and can also locate where an image has been altered. The basic m-sequence spatial approach [2] adds a modified m-sequence to the original image. Two types of sequences may be formed from an m-sequence: *unipolar* and *bipolar*. The elements of a bipolar sequence are {-1,1} and the elements of a unipolar sequence are {0,1}.

3. Two Dimensional Watermarks

This section overviews two two-dimensional watermarking techniques we present in [1]. The second technique provides better security and more precise localization for image alterations.

3.1 The Constant-W 2D Watermark

The *constant-W two-dimensional watermark* (CW2D) is formed as follows:

1. Form a $2^{16}-1$ period bipolar m-sequence.
2. Shape this to form a 256 x 256 watermark, W.

To mark the image, a 256 x 256 block of the image, X, is extracted

$$Y = X + W \quad (3)$$

where Y is the watermarked image block. This process is repeated until the entire image is marked. The total number of watermark blocks is image dependent. To verify a possibly forged image block Z, one must obtain the spatial crosscorrelation function:

$$R_{ZW}(\alpha,\beta) = \sum_x \sum_y Z(x,y)W(x-\alpha, y-\beta) \quad (4)$$

$$\delta = R_{YW}(0,0) - R_{ZW}(0,0) \quad (5)$$

The test statistic for the image block is δ. If $\delta < T$, where T is the test threshold, Z is genuine. If Z = Y, then $\delta = 0$. The advantages of this technique are presented in [1], as are the effects of linear and non-linear filtering.

3.2 Color Images

For 24 bit color images, each color plane may be treated as a monochrome image. If a color image is in the RGB color space, W can be added either to each color plane, or to one plane. For the YUV color space, one may choose to add W to the Y plane (luminance). Another possibility is to add W to the first color plane, then add an encrypted version, W_E, to the second plane, and finally an encrypted version of W_E to the third plane. Three-dimensional watermarks are also possible:

3.3 The Variable-W 2D Watermark

The *variable-W two-dimensional watermark* (VW2D) is generated by:

1. A bipolar m-sequence with a period of $2^{96}-1$ is obtained, and the first 128 bits are discarded.
2. The next 64 bits are shaped column-wise into an 8 x 8 block, W. The next 32 bits are discarded. This step repeats to form additional watermark blocks.
3. The marking and verification procedures are the same as in CW2D.

Advantages and disadvantages of VW2D are discussed in [1]. Also included are the effects of JPEG compression; the results indicate that VW2D works well with JPEG compression.

3.4 Random Bit Errors

We would like to investigate if VW2D can detect errors made to a JPEG compressed watermarked image. First, the original image X in Figure 1was watermarked with VW2D to form Y (Figure 2). The watermark, W_1, is shown in Figure 3. The watermarked image was JPEG compressed and decompressed to form Y_J. Then the LSB in Y_J was changed according to a given error rate to form an error image Z_J. This experiment was performed with both the unipolar and bipolar watermarks. To examine the effects for a wide range of data rates, four different quality factors were also used. Upon testing, the average of δ for all the blocks was obtained.

$$E\left[|\delta|\right] = \frac{1}{N_B}\sum_i\sum_j \delta_{ij} \qquad (6)$$

where δ_{ij} is the value of δ for the ith, jth block, and N_B is the number of 8 x 8 blocks in the image. Table 1 shows the values of E[|δ|] for each case.

Table 1. E[|δ|] after the addition of errors.

Q Factor:	55	65	75	85
Err. Rate	Bipolar Watermark			
0.005	0.265	0.264	0.265	0.267
0.0005	0.0277	0.0280	0.0282	0.0285
0.00005	0.00391	0.00391	0.00423	0.00423
	Unipolar Watermark			
0.005	0.138	0.138	0.138	0.138
0.0005	0.0122	0.0124	0.0124	0.0129
0.00005	0.00163	0.00163	0.00163	0.00163

The changes to the images in this experiment are minor. For an error rate of 0.005, only an average of 1966 pixels are changed in the entire 512 x 768 image. The fact that E[|δ|] is very small shows that our test statistic is not robust to small alterations. This is a potential problem. However, we could examine individual values of δ to see which blocks have been changed. We would also like to investigate the effects on E[|δ|] of compressing Z_J.

3.5 Random Bit Errors with Compression

This experiment is important in the following situation. An attacker could download a watermarked image that has been JPEG compressed at a known quality factor. The attacker would decompress the image and make changes to it. To store or transmit the forged image, the attacker would then re-compress the distorted image. We need to determine whether compressing the forged image Z_J will also produce a small variation in E[|δ|]. Z_J was first JPEG compressed and decompressed to form Z_C, using the same quality factor used for Y_J. The verification procedure was then performed. Table 2 lists the values of E[|δ|] for this experiment.

Table 2. E[|δ|] after adding errors + compression.

Q Factor:	55	65	75	85
Err. Rate	Bipolar Watermark			
0.005	0.349	0.267	0.150	0.105
0.0005	0.343	0.259	0.149	0.107
0.00005	0.343	0.259	0.149	0.108
	Unipolar Watermark			
0.005	1.05	0.726	0.429	0.265
0.0005	1.03	0.725	0.431	0.261
0.00005	1.03	0.725	0.431	0.263

Again E[|δ|] is quite small. A smaller quality factor will further increase E[|δ|], but not enough to make the changes easily detectable. These experiments show that very small changes to an image are very hard to detect by VW2D's authentication process. We will now discuss how the quantity E[|δ|] is used to verify images.

4. Interpreting the Test Statistic

We have made frequent mention of our test statistic, E[|δ|]. In this section we will develop a testing paradigm for using E[|δ|] to verify the authenticity of an image. We will discuss the range of E[|δ|] for an image to be fully authentic, authentic but forged, possibly authentic and completely inauthentic (or watermarked by a

different owner). To do this we will need two different thresholds for E[|δ|].

To develop these thresholds, we need to examine E[|δ|] in three situations:

1. When a watermark that is different from the embedded one is used to verify the image. This is equivalent to testing with a random watermark.
2. When the original unmarked image is tested with the watermark that will be embedded.
3. When an image has been watermarked a second time (IBM attack).

The proposed testing scenario is as follows. An image will be tested and E[|δ|] will be obtained. If E[|δ|] is greater than E[|δ|] obtained from testing with a different watermark, the image is considered either forged beyond recognition, or watermarked by a different owner. If E[|δ|] is less than E[|δ|] obtained from testing the original unmarked image, the test image is considered authentic. If E[|δ|] is between these values, then we can say that the image will likely belong to the watermark's owner, but has undergone some alterations. At this point, one would inspect the matrix, |δ|, to determine exactly where in the image the alterations took place.

Table 3 lists the values for E[|δ|] for six different 512 x 768 color images and four different testing and attack scenarios. Each image was watermarked (in the green plane only) with W_1. Figure 4 shows the original Tia image, and Figure 5 shows the watermarked version. Figure 6 has been watermarked twice. Figure 7 shows the watermark used for the Tia photo. All tests with W_1 produced E[|δ|] = 0, as expected. All tests used W_1 to obtain δ. The test scenarios are:

1. Testing with a different watermark: W_2, with an initial fill = 'abcdefghijkl'. When W_2 is substituted for W_1, E[|δ|] is very large – between 300 and 700.
2. Testing with a third watermark: W_3 with an initial fill = '0123456789AB'. Testing an image with W_3 also produces values of E[|δ|] between 300 and 700.
3. Testing the original: The watermarked image is imperceptibly different from the original unmarked image. The value of E[|δ|] should therefore be relatively small compared to 300. The actual values are between 4 and 9.
4. Preserving the original watermark given a re-watermarking attack: The IBM attack involves subtracting a second watermark from a watermarked image. To test our method against this type of attack, we marked the marked image with W_2, then tested it with W_1. The values of E[|δ|] are quite small. It also indicates that the image under test does indeed contain our watermark, W_1. This attack is discussed in more detail in Section 6.

Table 3. E[|δ|] for six different 768x512 images

Image	W_2	W_3	Orig.	W_1+W_2
Canyon	541.30	546.29	5.347	3.667
Tia	305.59	303.16	6.192	4.144
Vegetab.	461.53	466.54	4.047	3.107
Fruit	561.93	557.36	8.122	5.826
Glass	599.30	602.89	8.677	5.898
Money	647.36	647.48	7.060	5.890

A testing paradigm can now be established:

1. If E[|δ|] < 10, the image is perceptually identical to the original, watermarked image. It is fully authentic.
2. If 10 < E[|δ|] < 100, the image belongs to the owner of W_1, but has been changed.
3. If 100 < E[|δ|] < 200, the image probably belongs to the owner of W_1, but has been significantly altered – possibly beyond recognition.
4. If E[|δ|] > 200, the image has either been severely altered, or does not belong to the owner of W_1.

With this paradigm, all the images from Table 1 and Table 2 would be considered fully authentic. Also, all four sets of filtered images in [1] would pass.

5. Paletized Images

In many cases 24 bit color images are not used on web sites. Usually, 8 bit color images are used for simplicity. The common 8 bit color image format

is the Graphics Interchange Format (GIF) [10]. These images are paletized from 24 bits/pixel to 8 bits/pixel. Paletization (quantization) can be considered as a type of image attack; we therefore want to examine the performance of VW2D with paletized images. Paletization can be thought of as a nonlinear filtering process. It is also a lossy operation. This section examines the effects of paletizing a watermarked 24 bit RGB image. The procedure for marking a GIF image, X, is:

1. Convert X to a 24 bit color image, Y
2. Mark Y (green plane only)
3. Generate $R_{YW}(0,0)$
4. Quantize Y to 256 colors with a new color palette, call this Z
5. Distribute Z on the network

To test an image, Z, one must:

1. Convert Z to 24 bit RGB color.
2. Test the 24 bit version of Z in the usual way, and obtain E[|δ|].

The concern is that step 4 in the marking procedure, quantization, will destroy the watermark. Note: a new color palette must be generated in this step; if X's color palette is used, the watermark will be completely destroyed upon quantization of Y. The six images used previously were first converted from 24 bit RGB to 8 bit paletized images. These were then watermarked with W_1, and tested with the above procedure. Figure 8 is the 8 bit glass image; Figure 9 is the watermarked copy. Table 4 lists the values of E[|δ|] for each image.

Table 4. Results of paletization

Image	E[|δ|]
Canyon	4.278
Tia	2.676
Vegetables	6.436
Fruit	7.560
Glass	1.465
Money	1.366

The values of E[|δ|] are small and hence each quantized image would be fully authenticated. Preliminary tests with W_2 have produced E[|δ|] that is approximately 600.

6. The IBM Attack

An attack on a wide class of watermarking schemes, including the VW2D, is proposed in [11]:

$$Y_1 = X + W_1 \quad (7)$$

$$X_F = Y_1 - W_2 \quad (8)$$

$$\Rightarrow Y_1 = X_F + W_2 \quad (9)$$

X_F is known as the counterfeit original. The watermarked image, Y_1, now appears to be a marked version of X_F, marked with W_2. It cannot be discerned which is the actual original, X or X_F. Note that the counterfeit original was created *without* access to the true original, X.

One solution is to time stamp X when it is created [12]. Then, the owner with the earliest-dated original is the true owner. A watermark that is a non-invertible function of X would also securely determine ownership. Let H be the hash of X.

$$H = H(X) \quad (10)$$

In [11] it suggests using H to choose between two embedding equations. Our modification to VW2D is to use a time stamp, S, of X as the initial fill for watermark generation.

$$S = (H, T, U) \quad (11)$$

T is the time of X's creation, and U is the owner name.

$$Y_1 = X + W(S) \quad (12)$$

It is practically impossible to obtain the correct watermark, W(X), without knowing S. To successfully steal Y_1, one must find an X_F such that

$$X_F = Y_1 - W(S_F) \quad (8)$$

Since $W(S_F)$ depends on X_F, this task is very difficult.

7. Future Research

We plan to adapt the VW2D technique to watermark compressed and uncompressed video, including motion-JPEG, MPEG and H.263 sequences. Video watermarks require a fast verification procedure in order to authenticate video in real-time. The computation of δ may be too intensive for real-time video. For this reason we are researching alternative watermarking

algorithms which are more computationally efficient.

A postscript version of this paper is available via anonymous ftp to skynet.ecn.purdue.edu in the directory /pub/dist/delp/cisst97-secure. Sample images are available at the following web site: http://www.ece.purdue.edu /~ace

8. References

[1] Raymond B. Wolfgang and Edward J. Delp, "A watermark for digital images," *Proceedings of the 1996 International Conference on Image Processing*, Lausanne, Switzerland. Sept. 16-19, 1996, vol. 3, pp. 219-222.

[2] R. G. van Schyndel, A.Z.Tirkel, N.R.A Mee, C.F. Osborne, "A digital watermark," *Proceedings of the IEEE International Conference on Image Processing*, Austin, Texas, USA, November, 1994, vol. 2, pp. 86-90.

[3] R. G. van Schyndel, A. Z. Tirkel, C. F. Osborne, "Towards a robust digital watermark," *Proceedings of the ACCV-95 Conference*, Nanyang Technological University, Singapore, December 5-8, 1995.

[4] Ingemar J. Cox, Joe Kilian, Tom Leighton and Talal Shamoon, "Secure spread spectrum watermarking for images, audio and video," *Proceedings of the 1996 International Conference on Image Processing*, Lausanne, Switzerland, September, 16-19, 1996, vol. 3, pp. 243-246.

[5] J.-F. Delaigle, C. De Vleeschouwer, B. Macq, "Digital watermarking," *Proceedings of the IS&T/SPIE Conference on Optical Security and Counterfeit Deterrence Techniques*, San Jose, CA, USA, Feb. 1-2, 1996, vol. 2659, pp. 99-110.

[6] F. M. Boland, J. J. K. Ó Ruanaidh and C. Dautzenberg, "Watermarking digital images for copyright protection," *Proceedings of the International Conference on Image Processing and its Applications*, Edinburgh, Scotland, July 1995, pp. 321-326.

[7] Mitchell D. Swanson, Bin Zhu and Ahmed H. Tewfik, "Transparent robust image watermarking," *Proceedings of the 1996 International Conference on Image Processing*, Lausanne, Switzerland. Sept. 16-19, 1996, vol. 3, pp. 211-214.

[8] C. I. Podilchuk and W. Zeng. "Digital image watermarking using visual models," *Proceedings of the IS&T/SPIE Conference on Human Vision and Electronic Imaging II*, San Jose, CA, USA, Feb. 10-13, 1997, vol. 3016, pp. 100-111.

[9] Fred Mintzer, Albert Cazes, Francis Giordano, Jack Lee, Karen Magerlein and Fabio Schiattarella, "Capturing and preparing images of Vatican library manuscripts for access via Internet," *Proceedings of the IS&T 48th Annual Conference*, Washington, DC, USA, May, 1995, pp. 74 - 77.

[10] James D. Murray and William VanRyper, *Encyclopedia of Graphics File Formats*, Second Ed., O'Reilly & Associates, Inc. 1996, pp. 429-450.

[11] Scott Craver, Nasir Memon, Boon-Lock Yeo and Minerva Yeung. "Can invisible watermarks resolve rightful ownerships?", *Proceedings of the IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases V*, San Jose, CA, USA, Feb. 13-14, 1997, vol. 3022, pp. 310-321.

[12] Bruce Schneier, *Applied Cryptography*, Second Ed., Wiley & Sons, 1996.

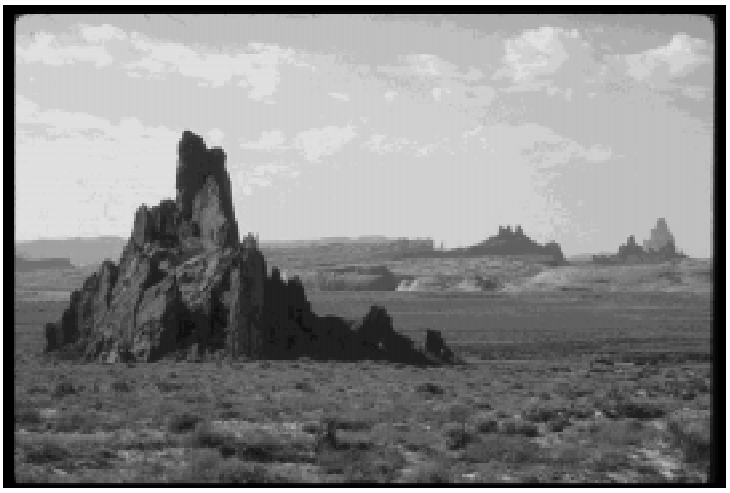

Figure 1. Original canyon image

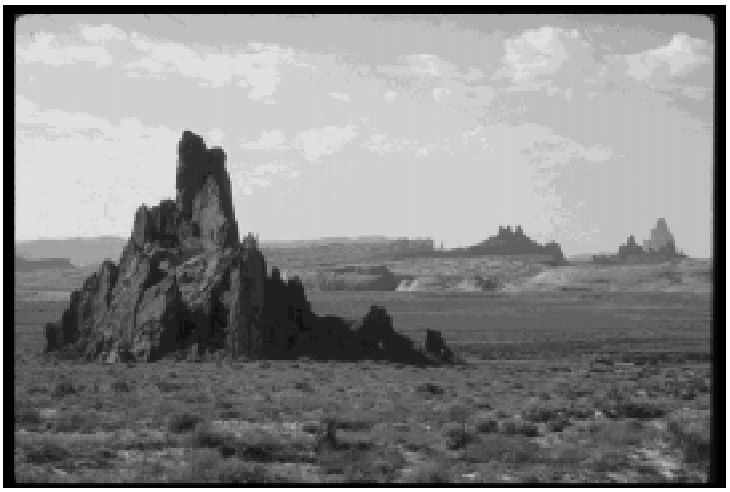

Figure 2. Watermarked canyon image

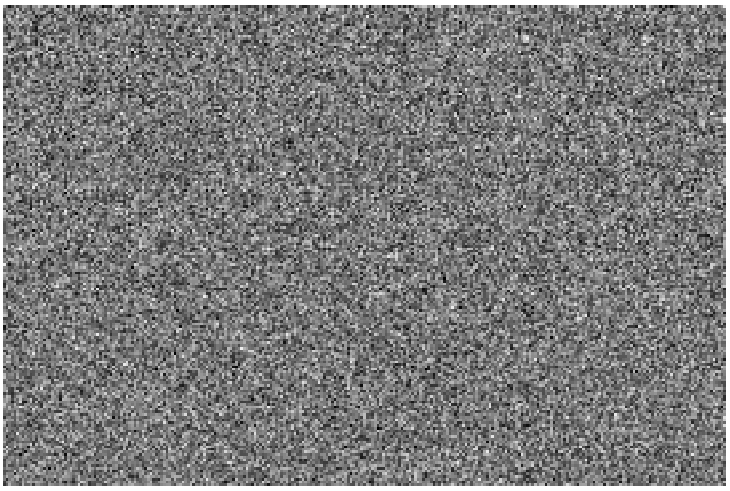

Figure 3. W_1 for canyon image

Figure 4. Original Tia picture

Figure 5. Watermarked picture of Tia

Figure 6. Tia watermarked with W_1 and W_2

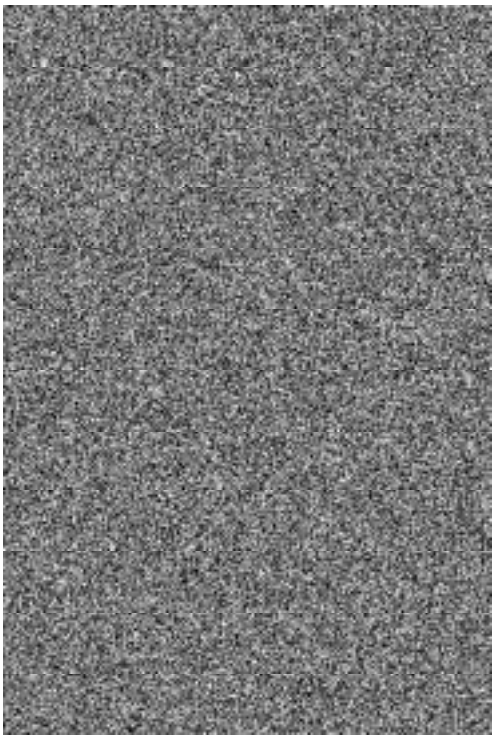

Figure 7. W_1 for Tia picture

Figure 8. Unwatermarked glass GIF image

Figure 9. Watermarked glass GIF image