CERIAS Tech Report 2001-138
A Review of Fragile Image Watermarks
by E Lin, E Delp
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

A Review of Fragile Image Watermarks

Eugene T. Lin and Edward J. Delp

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

1. ABSTRACT

Many image watermarks have been proposed to protect intellectual property in an age where digital images may be easily modified and perfectly reproduced. In a fragile marking system, a signal (watermark) is embedded within an image such that subsequent alterations to the watermarked image can be detected with high probability. The insertion of the watermark is perceptually invisible under normal human observation. These types of marks have found applicability in image authentication systems.

In this paper we discuss fragile marking systems and their desirable features, common methods of attack, and survey some recent marking systems.

1.1 Keywords

fragile watermarking, image authentication

2. Introduction

The age of digital multimedia has brought many advantages in the creation and distribution of image content but the ease of copying and editing also facilitates unauthorized use, misappropriation, and misrepresentation. Content providers are naturally concerned about these issues and watermarking, which is the act of embedding another signal (the watermark) into an image, have been proposed to protect an owners rights [1].

This work was supported a grant from Texas Instruments and an equipment grant from Intel. Address all correspondence to E. J. Delp, ace@ecn.purdue.edu, http://www.ece.purdue.edu/~ace, or +1 765 494 1740.

Many types of watermarks have been developed for a variety of applications. Watermarks may be visible or invisible, where a visible mark is easily detected by observation while an invisible mark is designed to be transparent to the observer and detected using signal processing techniques [1]. The process of embedding the watermark requires modifying the original image and in essence the watermarking process inserts a controlled amount of "distortion" in the image. The recovery of this distortion allows the one to identify the owner of the image. Invisible or transparent marks use the properties of the human visual system to minimise the perceptual distortion in the watermarked image[1] [2]. In the class of transparent watermarks one may further categorise techniques as robust or fragile. A robust mark is designed to resist attacks that attempt to remove or destroy the mark. Such attacks include lossy compression, filtering, and geometric scaling. A fragile mark is designed to detect slight changes to the watermarked image with high probability. The main application of fragile watermarks is in content authentication. Most of the work, as reported in the literature, in watermarking is in the area of robust techniques. Many important applications could benefit from the use of fragile watermarks.

3. Fragile Marking Applications

A fragile watermark is a mark that is readily altered or destroyed when the host image is modified through a linear or nonlinear transformation [3]. Fragile marks are not suited for enforcing copyright ownership of digital images; an attacker would attempt to destroy the embedded mark and fragile marks are, by definition, easily destroyed. The sensitivity of fragile marks to modification leads to their use in image authentication. That is, it may be of interest for parties to verify that an image has not been edited, damaged, or altered since it was marked.

Image authentication systems have applicability in law, commerce, defense, and journalism. Since digital images are easy to modify, a secure authentication system is useful in showing that no tampering has occurred during situations where the credibility of an image may be questioned. Common examples are the marking of images in a database to detect tampering [4][5], the use in a "trustwor-

thy camera" so news agencies can ensure an image is not fabricated or edited to falsify events [6], and the marking of images in commerce so a buyer can be assured that the images bought are authentic upon receipt [7]. Other situations include images used in courtroom evidence, journalistic photography, or images involved in espionage.

Another method to verify the authenticity of a digital work is the use of a signature system [8]. In a signature system, a digest of the data to be authenticated is obtained by the use of cryptographic hash functions [8][9]. The digest is then cryptographically signed to produce the signature that is bound to the original data. Later, a recipient verifies the signature by examining the digest of the (possibly modified) data and using a verification algorithm determines if the data is authentic. While the purpose of fragile watermarking and digital signature systems are similar, watermarking systems offer several advantages compared to signature systems [10] at the expense of requiring some modification (watermark insertion) of the image data. Since a watermark is embedded directly in the image data, no additional information is necessary for authenticity verification. (This is unlike digital signatures since the signature itself must be bound to the transmitted data.) Therefore the critical information needed in the authenticity testing process is discreetly hidden and more difficult to remove than a digital signature. Also, digital signature systems view an image as an arbitrary bit stream and do not exploit its unique structure. Therefore a signature system may be able to detect that an image had been modified but cannot characterise the alterations. Many watermarking systems can determine which areas of a marked image have been altered and which areas have not, as well as estimate the nature of the alterations.

3.1 Image Authentication Framework

The framework for embedding and detecting a fragile mark is similar to that of any watermarking system. An owner (or an independent third party authority) embeds the mark into an original image (see Figure 1).

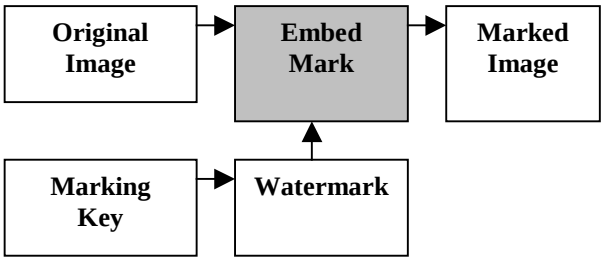

Figure 1: Watermark Embedding

The marking key is used to generate the watermark and is typically an identifier assigned to the owner or image. The original image is kept secret or may not even be available in some applications such as digital camera. The marked image may be transmitted, presented, or distributed. The marked image is perceptually identical to the original image under normal observation. See Figure 2 and Figure 3 for an example of original and marked images using the fragile marking technique described in [9][11].

Figure 2: Original Image

Figure 3: Watermarked Image

When a user receives an image, they use the detector to evaluate the authenticity of the received image (see Figure 4). The detection process also requires knowledge of "side information." This side information may be the marking key, the watermark, the original image, or other information. The detector is usually based on statistical detection theory whereby a test statistic is generated and from that test statistic the image is determined to be authentic. If it is not authentic then it would be desirable for the detector to determine where the image has been modified.

The side information used by the detector is very important in the overall use of a fragile watermark. Techniques that require that the detector have the original image are known as private watermarks while techniques that do require the detector to have the original image are known as public watermarks. To be effective a fragile watermarking system must be a public technique. In many applications the original image may never be available since it might have been watermarked immediately upon creation.

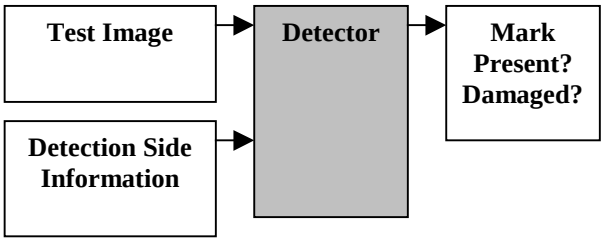

Figure 4: Watermark Detection

In database applications the owner or authority who marks the images is often the party interested in verifying that they have not been altered a subsequent time. For example, in a medical database it is important that any modifications to images be detected. In other applications, such as commerce, the verifying parties are distinct from the marking entity. In these cases, it is desirable to choose a system where the marking and detection information are distinct. In such a system, the ability to determine the authenticity of images does not also grant the ability to mark images. The vast majority of fragile systems described in the current literature do not implement this approach.

4. Features of Fragile Marking Systems

We now present desirable features of fragile marking systems, noting that the relative importance of these features will depend on the application. Applications may have requirements other than the ones mentioned. In addition to the features described below and the desired properties we previously mentioned, other properties can be found in [4][12][13]:

1. Detect tampering. A fragile marking system should detect (with high probability) any tampering in a marked image. This is the most fundamental property of a fragile mark and is a requirement to reliably test image authenticity. In many applications it is also desirable to provide an indication of how much alteration or damage has occurred and where it is located (see Feature 4 below).

2. Perceptual Transparency. An embedded watermark should not be visible under normal observation or interfere with the functionality of the image [1]. In most cases this refers to preserving the aesthetic qualities of an image, however if an application also performs other operations on marked images (such as feature extraction) then these operations must not be affected. Unfortunately there is not a lot of information how the "noise" introduced by marking process affects other image processing operations [14]. This is an open research problem. Also, transparency may be a subjective issue in certain applications and finding measures, which correlate well with perceived image quality, may be difficult.

3. Detection should not require the original image. This was discussed in detail in Section 3. As mentioned above the original image may nor exist or the owner may have good reason not to trust a third party with the original (since the party could then place their own mark on the original and claim it as their own.)

4. Detector should be able to locate and characterise alterations made to a marked image. This includes the ability to locate spatial regions within an altered image which are authentic or corrupt. The detector should also be able to estimate what kind of modification had occurred.

5. The watermark detectable after image cropping. In some applications, the ability for the mark to be detected after cropping may be desirable. For example, a party may be interested in portions (faces, people, etc.) of a larger, marked image. In other applications, this feature is not desired since cropping is treated as a modification.

6. The watermarks generated by different marking keys should be "orthogonal" during watermark detection. The mark embedded in an image generated by using a particular marking key must be detected only by providing the corresponding detection side information to the detector. All other side information provided to the detector should fail to detect the mark.

7. The marking key spaces should be large. This is to accommodate many users and to hinder the exhaustive search for a particular marking key even if hostile parties are somehow able to obtain both an unmarked and marked versions of a particular image.

8. The marking key should be difficult to deduce from the detection side information. This is particularly important in systems that have distinct marking and detection keys. Usually in such systems the marking key is kept private and the corresponding detection side information may be provided to other parties. If the other parties can deduce the marking key from the detection information then they may be able embed the owner's mark in images that the owner never intended to mark.

9. The insertion of a mark by unauthorised parties should be difficult. A particular attack mentioned in [4] is the removal of the watermark from a marked image and subsequently inserting it into another image.

10. The watermark should be capable of being embedded in the compressed domain. This is not the same as saying the watermark should survive compression, which can be viewed as an attack. The ability to insert the mark in the compressed domain has significant advantage in many applications.

5. Attacks on Fragile Marks

One must be mindful of potential attacks by malicious parties during the design and evaluation of marking systems. It may be practically impossible to design a system impervious to all forms of attack, and new methods to defeat marking systems will be invented in time. But certainly knowledge of common attack modes is a requirement for the design of improved systems.

The first type of attack is blind modification of a marked image (that is, arbitrarily changing the image assuming no mark is present). This form of attack should be readily recognized by any fragile mark, yet we mention it because it may be the most common type of attack that a marking system is to defeat. Variations of this attack include cropping and localized replacement (such as substituting one person's face with another.) The latter type of modification is a significant reason why an application may want to be able to indicate the damaged regions within an altered image.

Another type of attack is to attempt to modify the marked image itself without affecting the embedded mark or creating a new mark that the detector accepts as authentic. Some weak fragile marks easily detect random changes to an image but may fail to detect a carefully constructed modification. An example is a fragile mark embedded in the least-significant bit plane of an image. An attempt to modify the image without realizing that a mark is expressed in the LSB is very likely to disturb the mark and be detected. However, an attacker that may attempt to modify the image without disturbing any LSBs or substitute a new set of LSBs on a modified image that the detector classifies as authentic.

Attacks may also involve using a known valid mark from a marked image as the mark for another, arbitrary image [4]. The mark-transfer attack is easier if it is possible to deduce how a mark is inserted. This type of attack can also be performed on the same image; the mark is first removed, then the image is modified, and finally the mark is re-inserted.

An attacker may be interested in completely removing the mark and leaving no remnants of its existence (perhaps so they can deny ever bearing witness to an image which has their mark embedded in it). To do so, an attacker may attempt adding random noise to the image, using techniques designed to destroy marks (such as Stirmark [15]), or using statistical analysis or collusion to estimate the original image.

An attacker may also attempt the deduction of the marking key used to generate the mark. The marking key is intimately associated with an embedded mark, so if it is possible to isolate the mark the attacker can then study it in an attempt to deduce the key (or reduce the search space for the marking key). Once the key is deduced, the attacker can then forge the mark into any arbitrary image.

There are also known attacks that involve the authentication model itself and not so much on the specific mark in an image. Attacks on authentication systems over insecure channels are also discussed in [8] and similar vulnerabilities can apply to watermarking systems.

6. Examples of Fragile Marking Systems

We now survey some fragile marking systems described in the literature. We can classify the techniques as ones which work directly in the spatial domain or in the transform (DCT, wavelet) domains.

6.1 Spatial Domain Marks

Early fragile watermarking systems embedded the mark directly in the spatial domain of an image, such as techniques described in Walton [16] and van Schyndel et al. [17]. These techniques embed the mark in the least significant bit plane for perceptual transparency. Their significant disadvantages include the ease of bypassing the security they provide [3][18] and the inability to lossy compress the image without damaging the mark.

Wolfgang and Delp [11] extended van Schyndel's work to improve robustness and localization in their VW2D technique. The mark is embedded by adding a bipolar M-sequence in the spatial domain. Detection is via a modified correlation detector. For localization, a blocking structure is used during embedding and detection. This mark has been compared to other approaches using hash functions [9].

P. Wong describes another fragile marking technique in [19], which obtains a digest using a hash function. The image, image dimensions, and marking key are hashed during embedding and used to modify the least-significant bit plane of the original image. This is done in such a way that when the correct detection side information and unaltered marked image are provided to the detector, a bi-level image chosen by the owner (such as a company logo or insignia), is observed. This technique has localization properties and can identify regions of modified pixels within a marked image.

The technique of Yeung and Mintzer [3], whose security is examined in [10], is also one where the correct detection information results in a bi-level image. However, the embedding technique is more extensive than inserting a binary value into the least-significant bit plane. The marking key is used to generate several pseudo-random look-up tables (one for each channel or color component) that control how subsequent modification of the pixel data will occur. Then, after the insertion process is completed, a modified error diffusion process can be used to spread the effects of altering the pixels, making the mark more difficult to see. As discussed in [10], the security of the technique depends on the difficulty of inferring the look-up tables. The search space for the table entries can be drastically reduced if knowledge of the bi-level watermark image is known. A modification (position-dependent lookup tables) is proposed in [10] to dramatically increase the search space.

6.2 Transform Domain Marks

Various transformations, such as the discrete cosine transform (DCT) and wavelet transforms, are widely used for lossy image compression and much is known of how the actual transform coefficients may be altered (quantized) to minimize perceptual distortion [1]. There is also a great deal of interest in transform embedding for robust

image marking systems to make embedded marks more resilient to attacks.

There are advantages for fragile marking systems to use the transform domain as well. Many fragile marking systems are adapted from lossy compression systems (such as JPEG), which have the benefit that mark in embedded in the compressed representation. The properties of a transform can be used to characterize how an image has been damaged or altered. Also, applications may require a mark to possess some robustness to certain types of modification (such as brightness changes) yet be able to detect other modifications (e.g. local pixel replacemnet).

Wu and Liu [13] describe a technique based on a modified JPEG encoder. The watermark is inserted by changing the quantized DCT coefficients before entropy coding. A special lookup table of binary values (whose design is constrained to ensure mark invisibility) is used to partition the space of all possible DCT coefficient values into two sets. The two sets are then used to modify the image coefficients to encode a bi-level image (such as a logo.) To reduce the blocking effects of altering coefficients, it is suggested that the DC coefficient and any coefficients with low energy be not marked.

Kundur and Hatzinakos [12] and Xie and Arce [20] describe techniques based on the wavelet transform. Kundur embeds a mark by modifying the quantization process of Haar wavelet transform coefficients while Xie selectively inserts watermark bits by processing the image after it is in a compressed form using the SPIHT algorithm [21]. A wavelet decomposition of an image contains both frequency and spatial information about the image hence watermarks embedded in the wavelet domain have the advantage of being able to locate and characterize tampering of a marked image.

7. Conclusion

Fragile watermarking is the embedding of a signal (the watermark) into an image so that modifications to the resulting marked image can be detected with high probability. A fragile marking system is useful in a variety of image authentication applications. We feel that fragile watermarking has been somewhat ignored by the watermarking community in favor of robust techniques. There are many open research problems that need to be addressed in fragile watermarks such as the development of techniques that allow the detection of authenticity without permitting mark embedding. Many important applications can benefit from the use of fragile techniques

8. References

[1] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, "Perceptual Watermarks for Digital Images and Video", *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1108-1126, July 1999.

[2] M. Swanson, M. Kobayashi, A. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proceedings of the IEEE*, vol. 86, no. 6, pp. 1064-1087, June 1998.

[3] M. Yeung and F. Mintzer, "Invisible watermarking for image verification," *Journal of Electronic Imaging*, vol. 7, no. 3, pp. 578-591, July 1998.

[4] F. Mintzer, G. Braudaway, and M. Yeung, "Effective and ineffective digital watermarks," *Proceedings of the IEEE International Conference on Image Processing*, pp. 9-12, Santa Barbara, California, October 1997.

[5] F. Mintzer, G. Braudaway, and A. Bell, "Opportunities for watermarking standards," *Communications of the ACM*, vol. 41, no. 7, pp. 57-64, July 1998.

[6] G. Friedman, "The trustworthy digital camera: Restoring credibility to the photographic image," *IEEE Transactions on Consumer Electronics*, vol. 39, pp. 905-910, November 1993.

[7] P. W. Wong, "A public key watermark for image verification and authentication," *Proceedings of the IEEE International Conference on Image Processing*, vol. 1, pp. 455-459, Chicago, Illinois, October 1998.

[8] D. Stinson, *Cryptography Theory and Practice*, CRC Press, Boca Raton, 1995.

[9] R. Wolfgang and E. Delp, "Fragile watermarking using the VW2D watermark," *Proceedings of the IS&T/SPIE Conference on Security and Watermarking of Multimedia Contents*, pp. 204-213, San Jose, California, January 1999.

[10] N. Memon, S. Shende, and P. Wong, "On the security of the Yueng-Mintzer Authentication Watermark," *Final Program and Proceedings of the IS&T PICS 99*, pp. 301-306, Savanna, Georgia, April 1999.

[11] R. Wolfgang and E. Delp, "A watermark for digital images," *Proceedings of the IEEE International Conference on Image Processing*, vol. 3, pp. 219-222, 1996.

[12] D. Kundur and D. Hatzinakos, "Towards a telltale watermarking technique for tamper-proofing," *Proceedings of the IEEE International Conference on Image Processing*, vol. 2, pp. 409-413, Chicago, Illinois, October 1998.

[13] M. Wu and B. Liu, "Watermarking for image authentication," *Proceedings of the IEEE International Conference on Image Processing*, vol. 2, pp. 437-441, Chicago, Illinois, October 1998.

[14] S. Pankanti and M. Yeung, "Verification watermarks on fingerprint recognition and retrieval," *Proceedings of the IS&T/SPIE Conference on Security and Watermarking of Multimedia Contents*, pp. 66-78, San Jose, California, January 1999.

[15] Stirmark software: http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark, 1997.

[16] S. Walton, "Information authentication for a slippery new age," *Dr. Dobbs Journal*, vol. 20, no. 4, pp. 18-26, April 1995.

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

[18] J. Fridrich, "Image watermarking for tamper detection," *Proceedings of the IEEE International Conference on Image Processing*, vol. 2, pp. 404-408, Chicago, Illinois, October 1998.

[19] P. Wong, "A watermark for image integrity and ownership verification," *Final Program and Proceedings of the IS&T PICS 99*, pp. 374-379, Savanna, Georgia, April 1999.

[20] L. Xie and G. Arce, "Joint wavelet compression and authentication watermarking," *Proceedings of the IEEE International Conference on Image Processing*, vol. 2, pp. 427-431, Chicago, Illinois, October 1998.

[21] A. Said and W. Pearlman, "A new fast and efficient image codec based on set partitioning and hierarchical trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 3, pp. 243-250, June 1996.