A HYBRID WATERMARK FOR TAMPER DETECTION IN DIGITAL IMAGES

Jiri Fridrich
SUNY Binghamton and
Mission Research Corporation
Center for Intelligent Systems, Dept. of SS&IE, SUNY Binghamton, Binghamton, NY 13902-6000
fridrich@binghamton.edu

ABSTRACT

A new hybrid image authentication watermark obtained as a combination of a fragile and a robust watermark is presented. The fragile watermark has good localization and security properties but cannot be used to distinguish malicious changes, such as feature adding or removal, from innocent image processing operations. The hybrid watermark can be used to accurately pinpoint changes as well as distinguish malicious tamper from innocent operations. The authentication can be done without accessing any information about the original image. Examples of tamper detection on real imagery are given.

Powerful publicly available image processing software packages such as Adobe PhotoShop or PaintShop Pro make digital forgeries a reality. Feathered cropping enables replacing or adding features without causing detectable edges. It is also possible to carefully cut out portions of several images and combine them together while leaving barely detectable traces. Thus, the classical forensic techniques are not readily applicable to digital imagery.

1 INTRODUCTION

In the past, several different fragile watermarks have been introduced for authentication and tamper detection. One of the first techniques was based on inserting check-sums into the least significant bit (LSB) of image data. Walton [1] proposed a technique that uses a key-dependent pseudo-random walk on the image. The check-sum is obtained by summing the numbers determined by the 7 most significant bits and taking a remainder operation with a large integer N. The check-sum is inserted in a binary form in the LSB of selected pixels. The method is fast and on average modifies only half of the pixels by one gray level. Another example of a fragile watermark is the technique by Yeung and Wong [2,3] who use three key-dependent binary valued functions f_R, f_G, and f_B, $\{0, 1, \ldots, 255\} \rightarrow \{0,1\}$, to encode a binary logo B. All three color channels are perturbed in such a manner so that $B(i,j) = f_R(R(i,j)) \oplus f_G(G(i,j)) \oplus f_B(B(i,j))$. Error diffusion is employed to preserve the original colors. Delp et al. [4] describe two tamper detection techniques. The first one utilizes cryptographic hashes of small blocks that are stored separately from the image. Although this has good localization and security properties it is rather inconvenient to store the hashes separately. The second technique they describe is based on their δ statistics [5] introduced previously in connection with robust watermarking. By adjusting the decision threshold, it is possible to authenticate with a degree. However, since the watermarked image is necessary to calculate the statistics, the practical use of this technique is again questionable. A practical tamper detection method should not require access to the original or watermarked image. Fridrich and Goljan [6] describe a technique that can be used to *retrieve* the original content rather than just indicate which pixels or blocks have been tampered with. The idea is to self-embed the image into itself using a combination of lossy compression based on DCT transform of small image blocks and clever LSB encoding.

While fragile watermarks provide a very high probability of tamper detection they cannot distinguish between an innocent adjustment of brightness and replacing a person's face. Increasing the gray scales of all pixels by one would indicate a large extent of tampering, even though the image content has been unchanged for all practical purposes.

Zhu et al. [7] propose two techniques based on spatial and frequency masking. Their watermark is guaranteed to be perceptually invisible, yet it can detect errors up to one half of the maximal allowable change in each pixel or frequency bin depending on whether frequency or spatial masking is used. The image is divided into blocks and in each block a secret random signature (a pseudo-random sequence uniformly distributed in [0,1]) is multiplied by the masking values of that block. The resulting signal depends on the image block and is added to the original block quantized using the same masking values. Errors smaller than one half of the maximal allowable change are readily detected by this scheme. The error estimates are fairly accurate for small distortions. Kundur and Hatzinakos [8] propose a wavelet method that does not need to access the original image, and is not overly sensitive. Its main advantage is that it can localize the tamper in both the spatial and frequency domains. This is, however, achieved at the expense of precise localization of subtle changes on the level of individual pixels.

Fridrich [9,10] describes a technique capable of distinguishing malicious changes from innocent image operations or LSB shuffling. An image is divided into medium-size blocks and a robust spread-spectrum watermark is inserted into each block. If watermarks are present in all blocks with high probability, one can be fairly confident that the image has not been tampered with in any significant manner (such as adding or removing features comparable in size to the block). If the watermark correlation is lower uniformly over all image blocks, one can deduce that some image processing operation was most likely applied. If one or more blocks show very low evidence for watermark presence while other blocks exhibit values well above the threshold, one can estimate the probability of tampering and with a high probability decide whether or not the image has been tampered with

It appears that no single scheme can have both precise localization properties without being too sensitive. Indeed these two requirements are in conflict. On the other hand, it should be possible to combine a robust watermark with a fragile one, if the fragile watermark is embedded as the second one. The fragile watermark is usually very weak and should not influence the robust one in any significant manner. This hybrid watermark can, therefore, enjoy the good properties of both watermarks. If a subtle change is made to a highly localized group of pixels, such as changing the eyes color in a portrait photograph, the fragile watermark can be used to precisely localize the change. On the other hand, a simple lossy compression or applying a filter to the image will be indicated as non-malicious tamper because the robust watermark will survive. In this paper, we combine the fragile watermark as described in [2] with a semi-robust watermark on medium size blocks developed in our previous work [9,10].

2 HYBRID WATERMARK

The hybrid watermark consists of two watermark patterns embedded consecutively in the image. First, the robust watermark is embedded and then the fragile watermark is embedded on top of the robust one. Since the fragile watermark constitutes a small distortion, the robust watermark should be influenced only insignificantly.

2.1 Robust watermark

The details of this technique are described in [9,10] and are omitted in this paper for brevity. The technique starts with dividing the image into medium size blocks of 64×64 pixels. The size of the block is a compromise between the robustness to image distortion and the size of the smallest detectable feature tampering. Each block is DCT transformed, the coefficients are ordered in a zigzag manner as in JPEG compression, and the middle 30% of the coefficients D_k are modulated by a Gaussian signal S_k with zero mean and unit standard deviation by simply adding both signals

$$D'_k = D_k + \alpha S_k, \quad k = 1, \ldots, N_m,$$

where D'_k denotes the modulated DCT coefficients, α is the watermark strength, and N_m (0.3×64^2) is the number of modified coefficients. The watermarked block is obtained by performing the inverse DCT using the modulated coefficients D'_k . In order to make the watermarking scheme secure when a large number of images need to be watermarked, such as in digital cameras or for watermarking video, the watermark should be a strong function of the image block and a secret key. In [8,9,11] it is explained how a Gaussian signal can be synthesized so that it is a continuous function of the image block but at the same time is depends sensitively on a secret key.

The detection of the watermark again proceeds by blocks. Each block is DCT transformed, and the spread spectrum signal is synthesized from the tampered block using the secret key. The watermark presence is evaluated using the probability that the decoded watermark would partially agree with the original one by pure chance. The logarithm at the base 10 of this probability is used in Figures 2–5.

2.2 The fragile watermark

As a fragile watermark for our technique, we used the approach of Yeung and Wong [2,3]. For a gray-scale image, a key-dependent binary valued function

$$f: \{0, 1, \ldots, 255\} \rightarrow \{0,1\},$$

is used to encode a binary logo B. The gray scales are perturbed in such a manner so that $B(i,j) = f_R(R(i,j))$. Error diffusion is employed to preserve the original colors. The image authenticity is easily verified by checking the relationship $B(i,j) = f_R(R(i,j))$ for each pixel (i,j).

3 EXPERIMENTAL RESULTS

We tested the performance of the hybrid watermark by introducing a wide variety of image distortions and combining intentional tampering with common image processing operations. In Figure 1, you can see the test image "Lenna" with both watermarks. The average power of the robust watermark was 10.07 and its amplitude was confined to the interval [−13,13], while the average power of the fragile watermark was 1.13 in [−4,5]. The image in Figure 1 also shows the structure of 64×64 blocks with block numbers superimposed on it.

In our experiments, we used the popular image processing tool PaintShop Pro [12]. Our first experiment consisted of softening Lenna's eyes with the retouch tool. The results of tamper detection are shown in Figure 2. As expected, the robust watermark does not indicate any tamper, while the fragile watermark clearly localizes the areas where changes have occurred. In the second experiment, we have adjusted the brightness of the watermarked image by 10% and saved at the default compression level in PaintShop Pro (roughly 10 times smaller image). While the fragile watermark indicated that the whole image has been tampered with, the robust watermark gives us sufficient evidence that the lady in the picture is Lenna and that no major changes have occurred (see Figure 3). In the third experiment, we replaced Lenna's face with the face of the American actress Meg Ryan (see Figure 4). Both the fragile watermark and the robust watermark pinpointed the tampered areas. The robust watermark indicated tamper only for block No. 11, because a large portion of this block has been replaced. The remaining blocks have been tampered only partially, which shows up as high evidence for watermark presence. A pirate who created the forgery shown in Figure 4 might want to soften the image little bit in order to cover the lines caused by cutting and pasting. He does so by applying the softening filter (a low pass filter) to the whole image. The result is shown in Figure 5. After applying the detection procedure to the tampered image, the fragile watermark indicates that the whole image has been tampered with. This is due to the softening filter. The robust watermark, however, clearly shows that most of the image blocks are essentially untampered, while blocks No. 7, 11, and 14 have been tampered in a significant manner, and blocks No. 6 and 10 also suffered a lot of distortion. The essential content of the remaining blocks remains untouched.

Figure 1. The watermarked image "Lenna" with outlined blocks and block numbers.

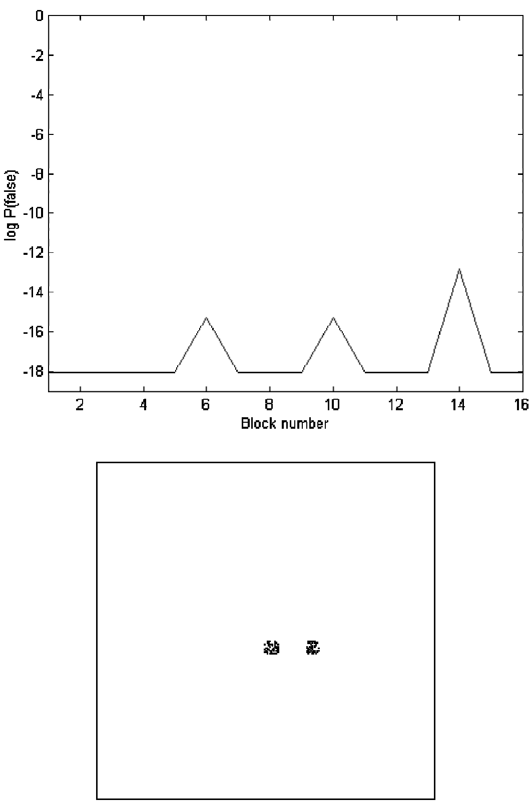

Figure 2. (Retouched eyes) Presence of the robust watermark (above); Fragile watermark indicated tampered areas with black dots (below).

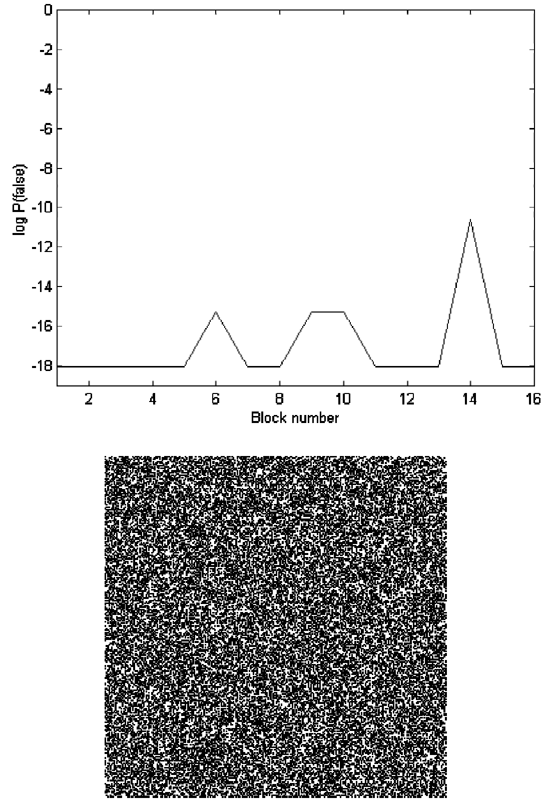

Figure 3. (After brightness adjustment and JPG compression) Presence of the robust watermark (above); Fragile watermark indicated tampered areas with black dots (below).

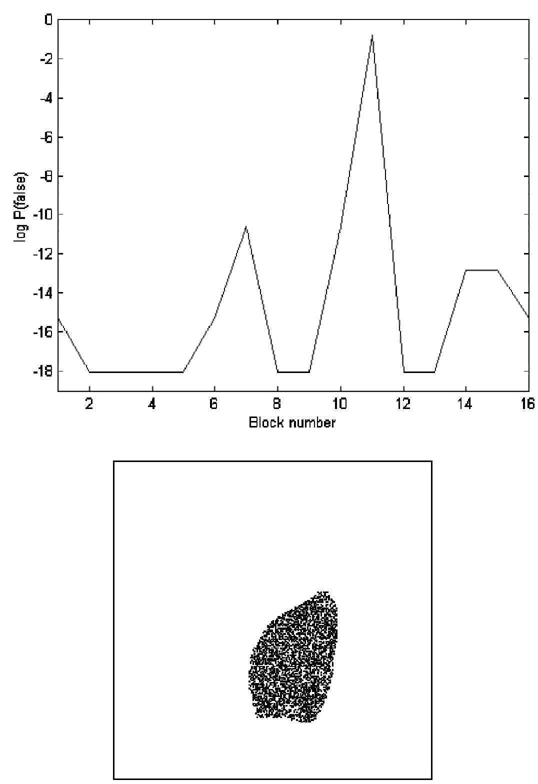

Figure 4. (Replaced face and softened) Presence of the robust watermark (above); Fragile watermark indicated tampered areas with black dots (below).

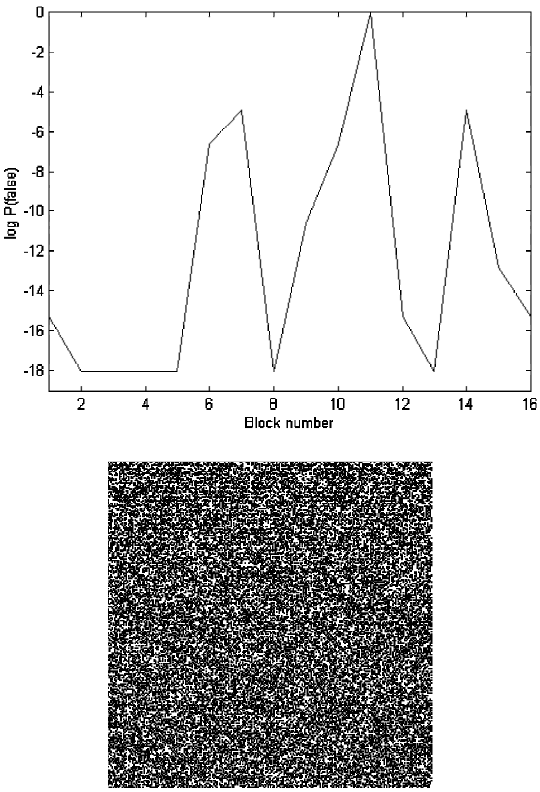

Figure 5. Replaced face. Presence of the robust watermark (above); Fragile watermark indicated tampered areas with black dots (below).

Figure 6. Lenna's face replaced with Meg Ryan's face.

4 SUMMARY

In this paper, we propose to combine fragile watermarks with robust watermarks on medium size blocks. The resulting hybrid watermark enjoys properties of both, which makes it a valuable tool for authentication of images and detection of tamper. The robust watermark is capable of distinguishing malicious changes, such as feature replacement/adding, from innocuous changes due to common image processing operations, such as filtering, contrast/brightness adjustment, or lossy compression. The performance of the hybrid watermark has been tested in a series of different combinations of intentional temper and common image processing operations.

Our future effort will be focused on development of image authentication tools that are can survive some simple geometrical operations, such as resizing or cropping. This could be achieved by using a different combination of watermarking schemes or using alternative watermarking techniques that encode the watermark into the mutual relationship between color channels.

5 ACKNOWLEDGEMENTS

The work on this paper was supported by Air Force Research Laboratory, Air Force Material Command, USAF, under a Phase II SBIR grant number F30602-98-C-0049. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there

REFERENCES

[1] S. Walton, “Information Authentication for a Slippery New Age”, *Dr. Dobbs Journal*, vol. 20, no. 4, pp. 18–26, Apr 1995.

[2] M. Yeung, and F. Mintzer, "An Invisible Watermarking Technique for Image Verification", *Proc. ICIP'97*, Santa Barbara, California, 1997.

[3] P. Wong, "A Watermark for Image Integrity and Ownership Verification", *Proc. IS&T PIC Conference*, Portland, Oregon, 1998.

[4] R. B. Wolfgang and E. J. Delp, "Fragile Watermarking Using the VW2D Watermark", *Proc. SPIE, Security and Watermarking of Multimedia Contents*, San Jose, California, Jan 25–27, 1999, pp. 204–213.

[5] R. B. Wolfgang and E. J. Delp, “A Watermark for Digital Images”, *Proc. IEEE Int. Conf. on Image Processing*, vol. 3, pp. 219–222, 1996.

[6] J. Fridrich and M. Goljan “Protection of Digital Images Using Self Embedding”, *Symposium on Content Security and Data Hiding in Digital Media*, New Jersey Institute of Technology, May 14, 1999.

[7] B. Zhu, M. D. Swanson, and A. Tewfik, “Transparent Robust Authentication and Distortion Measurement Technique for Images”, preprint, 1997.

[8] D. Kundur and D. Hatzinakos, "Towards a Telltale Watermarking Technique for Tamper Proofing", *Proc. ICIP*, Chicago, Illinois, Oct 4–7, 1998, vol 2.

[9] J. Fridrich, “Image Watermarking for Tamper Detection”, *Proc. ICIP ’98*, Chicago, Oct 1998.

[10] J. Fridrich, “Methods for Detecting Changes in Digital images”, ISPACS, Melbourne, November 4–6, 1998.

[11] J. Fridrich, “Robust Bit Extraction From Images”, *ICMCS'99*, Florence, Italy.

[12] PaintShop Pro, http://www.jasc.com.