

Robust digital watermarking based on key-dependent basis functions

Jiri Fridrich¹
Center for Intelligent Systems
SUNY Binghamton
Binghamton, NY 13902-6000
e-mail: fridrich@binghamton.edu

² Lt Arnold C. Baldoza and Richard J. Simard
Air Force Research Laboratory/IFEC
32 Hangar Road
Rome, NY 13441-4114
e-mail: {baldozaa, simardr}@rl.af.mil

Abstract: In this paper, we introduce the concept of key-dependent basis functions and discuss its applications to secure robust watermarking for copyright protection and to designing secure public black-box watermark detectors. The new schemes overcome a possible security weakness of global, non-adaptive schemes that apply watermark patterns spanned by a small number of publicly known basis functions. The watermark is embedded into the projections of an image onto the secret set of key-dependent functions (patterns). The robustness of the watermarking scheme with respect to filtering, lossy compression, and combinations of many other attacks is studied. Finally, we propose a candidate for a watermarking scheme that enables the construction of a secure public watermark detector.

Keywords: Robust image watermarking, attacks, orthogonal patterns, secure public watermark detector

1. Introduction

Digital images and digital video-streams can be easily copied. Even though such copying may violate copyright laws, it is widespread. The ease with which electronic images may be copied without loss of content significantly contributes to illegal copying. One of the goals of digital watermarks is authentication for copyright protection. To prove the ownership of an image, a perceptually invisible pattern (a watermark) is embedded into the image and ideally stays in the image as long as the image is recognizable. This means that the watermark must be embedded in a robust way and withstand any attempts to remove it using image processing tools as well as a targeted intentional removal based on the full knowledge of the watermarking scheme.

As pointed out by Cox et al. [1] and Miller [2] the watermark should be embedded in the most perceptually important features in the image, otherwise it would be too sensitive to compression schemes capable of removing *redundant* information. While schemes that adapt the watermark strength according to local image properties provide higher robustness [3–5], it is not entirely clear whether or not they provide higher degree of security because the watermarked image provides a clue about the strength and location of the watermark. This may be a handicap if such schemes were to be combined with public watermark detectors [11,12]. In addition to that, future compression schemes capable of removing *irrelevant* information may disrupt such watermarks [6–8].

¹ The author has been supported by an SBIR research grant “Intelligent C4I Technologies”.

Schemes that embed watermarks into the projections onto smooth orthogonal basis functions such as, discrete cosines, are typically very robust and less sensitive to synchronization errors due to skipping of rows of pixels, and/or permuting of nearby pixels than techniques that embed watermarks using pseudo-noise patterns [9,10]. However, if the watermark pattern is spanned by a relatively small number of publicly known functions, it may be possible to remove the watermark or disrupt it beyond reliable detection if a portion of the watermark pattern can be guessed or is known², or when the embedding key becomes partially available. The plausibility of such an attack is demonstrated in Section 2, where a simplified NEC scheme is analyzed.

This observation lead us towards investigating general, key-dependent orthogonal basis functions as a replacement for publicly known bases, such as discrete cosines. We believe that such techniques may significantly increase the security of watermarking schemes. A technique that utilizes key-dependent orthogonal functions (patterns) is described and analyzed in Section 3. Another important motivation for this paper was the problem of designing a secure black-box public watermark detector. In Section 4, we describe a candidate for such a secure detector based on key-dependent bases. Secure public detectors find important applications in copy-control of Digital Video Disks [11–15]. In Section 5, we summarize the paper and outline future research directions.

2. An attack on global watermarking schemes

It is important that a partial knowledge of the watermark should not enable a pirate to remove the entire watermark or disturb it beyond reliable detection. Below, we show that it is indeed possible in certain cases to reconstruct the watermark pattern based on the assumption that the watermark becomes known in some small area. This assumption is not that unreasonable as it may seem at first. For example, one can make a guess that certain portion of the original image had pixels of uniform brightness or of a uniform gradient, or an attacker may be able to foist a piece of his image into a collage created by somebody else. If this is the case, then the knowledge of a portion of the watermark pattern may give us additional constraints to disturb or eliminate the whole watermark. This is especially relevant for watermark patterns spanned by publicly known functions. Below, we describe an attack that can be applied to any non-adaptive robust watermarking technique, invertible or not, if some portion of the original unwatermarked image is known or can be guessed, and if the watermark is mostly spanned by some small number of Fourier modes. The attack attempts to find the coefficients of the lowest frequency DCT coefficients based on the “known” pixel values. A set of linear equations completed with a stabilizing functional makes the inversion possible.

In the watermarking technique proposed by Cox et al. [1], the watermark is embedded into a selected set of discrete cosine coefficients (the highest energy 1000 frequency coefficients). The logic behind this technique is to hide the watermark into the most

² This can happen in a collage consisting of several images.

perceptive modes of the image to achieve a high degree of robustness with respect to lossy compression and most common image processing techniques. The watermark is spanned by 1000 highest frequency discrete cosines. The non-locality of the watermark pattern could be potentially dangerous if an attacker is able to guess the original, unwatermarked values of some pixels. What makes the attack hard to mount, however, is the fact that discrete cosines are not linearly independent on proper subsets of the image, and, depending on the number of discrete cosines spanning the watermark, we may not have enough constraints to exactly recover the whole watermark.

In order to demonstrate the plausibility of the proposed attack, we performed the following experiment with a weakened version of the scheme proposed in [1]. The watermark is embedded into the lowest 50 coefficients v_k of the DCT according to the formula

$$v_k' = v_k (1 + \alpha \eta_k),$$

where v_k' are the modified DCT coefficients, η_k is a Gaussian sequence with zero mean and unit variance, and α is the watermark strength (also related to watermark's visibility). The watermarked image is obtained by applying the inverse DCT to the coefficients v_k' . In our experiments, we took $\alpha = 0.1$.

Let us assume that there is a region containing P pixels (i, j) in the image for which the original pixel values are known. Using the inverse DCT transformation, we can express the difference, $I_w - I$, between the watermarked and the original image as

$$(I_w - I)(i, j) = \frac{2}{\sqrt{M \times N}} \sum_{k=1}^J c_1(r_k) c_2(s_k) \cdot \alpha \cdot \eta(k) \cdot V(r_k, s_k) \cos \frac{\pi}{2M} r_k (2i + 1) \cos \frac{\pi}{2N} s_k (2j + 1)$$

where

$$c_1(r) = 1/\sqrt{2} \text{ when } r = 0 \text{ and } c_1(r) = 1 \text{ otherwise}$$

$$c_2(s) = 1/\sqrt{2} \text{ when } s = 0 \text{ and } c_2(s) = 1 \text{ otherwise}$$

and $V(r, s)$ denotes the coefficient matrix of DCT. The indices (r_k, s_k) , $k = 1, \dots, J$ correspond to the 50 lowest frequency discrete cosines that have been modified. The above equation describes a linear system of P equations for J unknowns $\eta_k \cdot V(r_k, s_k)$. Since our goal is to obtain the sequence η_k , we need to use the DCT of the watermarked image to calculate η_k . This can be done easily because the DCT of I_w gives us $V(r_k, s_k) (1 + \alpha \eta_k)$. The number of equations is determined by the number, P , of pixels (i, j) for which the original gray levels can be estimated or are known. Even though the number of pixels, P , may exceed J , the rank of the matrix may be smaller than J because discrete cosines do not generally form a set of linearly independent functions on proper subsets of the image.

In our experiment, we used a test image containing 128×128 pixels with 256 gray levels. The image has a small area of pixels in the upper right corner that has a constant

luminance of 192 (see Figure 1). We took $P = 862$ pixels that had constant brightness in the original unwatermarked image. Then, a watermark was inserted into the lowest $J = 50$ coefficients of the DCT using the algorithm above. The resulting overdetermined system of equations was solved for η_k . The original and recovered watermark sequences are shown in Figure 2. The watermark has been recovered almost exactly. It was not recovered completely accurately because the matrix of the system of equations was ill conditioned.

By increasing the number of modified coefficients in watermark embedding, this attack becomes harder to perform because the rank of the matrix is basically determined by the number of pixels, P , their spatial arrangement, and the image size. By increasing the number of modified coefficients, J , to 100, the MATLAB linear solver could not recover the watermark sequence due to an ill-conditioned matrix. It is not surprising that a general linear system solver breaks down in such cases. More sophisticated techniques that were not investigated so far could be put to work. For example, one could add constraints that will make the problem of finding the watermark sequence better conditioned. One obvious possibility is to use stabilizing functionals that would give penalty to sequences that do not satisfy Gaussian statistics. Even though such methods are usually computationally expensive, speed is obviously not a critical issue in watermark breaking.

The above attack can be mounted against any non-adaptive watermarking technique that inserts watermarks by modifying a relatively small set of selected coefficients in the DCT or other publicly known image transformation. The attack can be thwarted by using a larger number of coefficients in those transforms, or by adapting the watermark to the image content. As argued in the introduction and in [2], global schemes that embed watermarks into projections on orthogonal basis functions may have certain advantages over adaptive techniques. In the next section, we investigate watermarking techniques in which the orthogonal basis of discrete cosines is replaced by a set of general, random, smooth, orthogonal patterns that sensitively depend on a secret key.



Figure 1 A test image with a small area of pixels of constant brightness (the upper right corner)

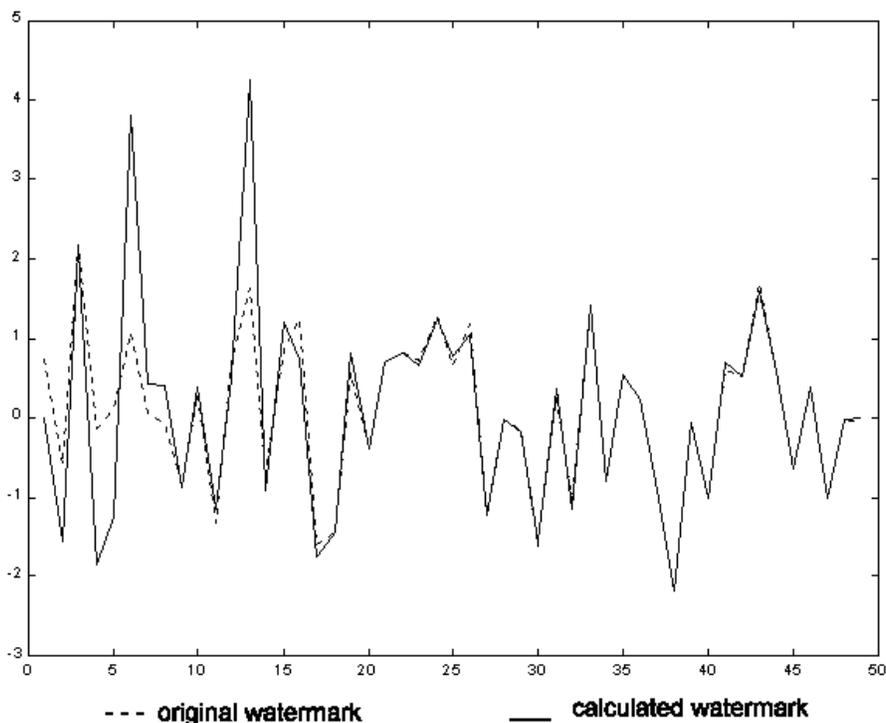


Figure 2 Comparison of the original and the recovered watermark of length 50.

3. Orthogonal patterns and their use in digital watermarking

The high robustness of the method of Cox et al. [1] is due to the fact that the watermark is placed into the most perceptive Fourier modes of the image. As argued above, the fact that a publicly known transformation is used can potentially become dangerous if portions of the original unwatermarked image can be guessed. The security of the scheme and its versatility could be increased if a different set of orthogonal basis functions would be used depending on a secret key. Since the basis functions or, equivalently, the key for their generation, will be kept secret, the watermark pattern could be spanned by a smaller number of functions thus enabling us to embed more bits in a robust and secure manner. To achieve this goal, we need a method for generating a set of orthogonal random functions (patterns) that sensitively depend on each bit of a secret key and possibly on an image hash. To guarantee good robustness properties, the generated patterns should have their energy concentrated mainly in low frequencies.

It is not necessary to generate a *complete* set of orthogonal basis functions since only a relatively small number of them is needed to span a watermark pattern. One can calculate projections³ of the original image onto a set of J orthogonal functions, and modify the projections so that some secret information is encoded. Let us denote such functions f_i ,

³ The dot product of two images A_{ij} and B_{ij} is defined as $\langle A, B \rangle = \sum_{i=1}^M \sum_{j=1}^N A_{ij} B_{ij}$

$i = 1, \dots, J$. Assuming that the functions are orthogonal to each other, the system of J functions can be completed by $MN-J$ functions g_i to a complete orthogonal system. The original image I can then be written as

$$I = \sum_{i=1}^J c_i f_i + g, \quad c_i = \langle f_i, I \rangle,$$

where g is a linear combination of functions g_i that are orthogonal to f_i . The watermarking process is realized by modifying the coefficients c_i . Furthermore, the watermarked image I_w can be expressed as

$$I_w = \sum_{i=1}^J c'_i f_i + g, \quad c'_i = \langle f_i, I_w \rangle = (1 + \alpha w_i) c_i,$$

where c'_i are the modified coefficients, α determines watermark's strength and visibility, and w_i is a watermark sequence. Given a modified watermarked image Im ,

$$Im = \sum_{i=1}^J c''_i f_i + g', \quad c''_i = \langle f_i, Im \rangle,$$

we can calculate the modified coefficients by evaluating the projections of Im onto the functions f_i . A cross-correlation $corr$ of the differences $c''-c'$ with $c'-c$,

$$corr = \frac{(c''-c')(c'-c)}{\|c''-c'\| \|c'-c\|}, \quad (1)$$

is compared to a threshold to decide about the presence of a watermark.

If the watermarking method is used for copyright protection, the sequence w_i should depend on the image hash in order to prevent a forgery of the original image [16]. The orthogonal patterns do not have to be image dependent because they depend on some initial secret key and it would be clearly computationally infeasible to forge them given a certain watermark sequence w_i . In Section 4, we describe a modification of this scheme in which the original image is not needed for watermark extraction. In this case, the watermark sequence does not have to be a part of the secret key and can carry several bits of useful information.

For practical implementation, we need a method for generating a set of random, smooth, orthogonal patterns whose power is concentrated in low frequencies. The patterns should sensitively depend on each bit of the secret key. There is obviously more than one way to achieve this task. One possibility would be to use some known orthogonal basis, such as the discrete cosines, and build a new basis from them. For example, one could choose M lowest frequency discrete cosines, randomly divide them into M/G groups of G cosines, and linearly combine the functions in each group to get M/G random, smooth, orthogonal patterns. This method is equivalent to embedding watermark patterns into linear combinations of selected G -tuples of DCT coefficients. Although this approach does not produce patterns that are "truly random", it has low computational complexity and can be easily implemented. Detailed investigation of this approach will be a part of future research.

In this paper, we opted for a general approach that does not use any orthogonal basis as the building blocks. We generate a set of J pseudo-random black and white patterns using a cryptographically strong pseudo-random number generator seeded with the secret key.

The patterns are further smoothed by applying a low-pass filter to them. To make the patterns orthogonal, the Gram-Schmidt orthogonalization procedure is applied. Finally, the functions are normalized to obtain an orthonormal system. This way, we obtain a set of J orthonormal functions that have their power concentrated in low frequencies. Moreover, the patterns sensitively depend on each bit of the secret bit-string. Although this approach is rather computationally expensive, we feel that it is important to investigate the properties of this most general scheme in order to prove the viability of the concept as a whole.

The scheme for embedding watermarks can be described as follows: Secret key \rightarrow (pseudo-random number generator + smoothing) \rightarrow a set of J random, smooth patterns \rightarrow (Gram-Schmidt orthogonalization process) \rightarrow a set of J orthonormal, random, smooth patterns \rightarrow (modifying projections according to some key {and image hash}) \rightarrow watermarked image. In our scheme, the first coefficient plays the role similar to the DC term in a DCT. To preserve the energy of the watermarked image, the first coefficient c_1 is left unmodified.

To retrieve the watermark, we calculate the projections c_i onto the J secret functions f_i . The projections c_i are then compared with those of the watermarked image and the original unwatermarked image by calculating the correlation (1). Based on the value of this correlation, we decide whether or not a watermark is present. To avoid large memory requirements to store all orthogonal patterns, only the image hash and the secret key need to be stored. The orthogonal patterns can be generated for each detection attempt.

A pseudo-code for the watermarking algorithm (gray scale $N \times N$ images):

```
begin_algorithm
read image I;           // I is a matrix of integers 0, ..., 255
convert I to an intensity matrix X; //  $x_{ij} \in [0,1]$ 
seed=secret_bitstring; // Initialize a PRNG with a secret bit string
```

Step 1 (Generate J pseudo-random binary patterns and smooth them)

```
for k=1 to J
    using a PRNG, generate an  $N \times N$  binary pattern  $Z^k = Z^k_{ij}$ ,  $1 \leq i, j \leq N$ ;
     $Z^k = \text{smooth}(Z^k)$ ;
end_for
```

Step 2 (Orthogonalize the smoothed patterns using Gram-Schmidt orthogonalization procedure)

```
for k=1 to J

$$Z^k = Z^k - \sum_{s=1}^{k-1} \langle Z^s, Z^k \rangle Z^s$$


$$Z^k = \frac{Z^k}{\|Z^k\|}$$

end_for
```

Step 3 (Calculate the projections and modify them to embed a watermark)

for $k=1$ to J

$$c_k = \langle Z^k, X \rangle$$

$$c_k' = c_k (1 + \alpha w_k)$$

end_for

Step 4 (Calculate the watermarked image X_w)

$$X_w = X + \alpha \sum_{j=2}^J w_j c_j f_j$$

Convert X_w to a gray-scale image I_w ;

end_algorithm

The coefficient α determines the visibility of the watermark and its robustness. The watermarking scheme could be applied either globally to the whole image, or locally. In the global scheme, the support of the functions f_i is the whole image. This makes the scheme computationally very expensive with large memory requirements. For an $N \times N$ grayscale image, one needs JN^2 bytes to store all J orthogonal patterns. This number could become prohibitively large with even for moderate values of N (such as $N = 256$). The most time consuming part of the algorithm is the Gram-Schmidt orthogonalization procedure. Its computational complexity is $O(J^2N^2)$. Thus, the choice of the number of orthogonal patterns, J , turns out to be critical. If J is chosen too small, the correlation used for detection of watermarks can have occasionally large values due to random correlations. If J is chosen too large (of the order of 1000 or larger), the computational complexity of the scheme becomes unreasonably large. A good compromise is to break the image into smaller subregions that are watermarked separately using different sets of orthogonal patterns and average the correlations from multiple subregions. The averaging will decrease the values of random correlations, while keeping the robustness sufficiently high and at reasonable computational requirements.

The combination of the following parameters is crucial to obtaining a computationally effective and robust watermarking scheme: (i) size of the subregions, (ii) watermark length J , (iii) watermark strength α . A detailed study of how the performance of the new scheme is influenced by different combinations of these parameters is necessary and will be a part of the future research. The tests that were performed so far indicate that the new scheme is very robust with respect to blind attempts to remove the watermark. It also provides higher degree of security when compared to global schemes that form the watermark from publicly known basis functions, due to the fact that the orthogonal patterns are kept secret and are generated from a secret key. Both the global and local versions of the new watermarking scheme were implemented as m-functions in Matlab and tested for robustness. Some preliminary results are included below.

Test 1: Global scheme

Even though the global scheme is not suitable for practical use due to the immense computational and memory requirements, we nevertheless performed tests on a 64×64 image (see Figure 3).



Figure 3 Original image.



Figure 4 Watermarked image.

The image was watermarked using $J = 100$ orthogonal patterns, and tested for presence of 100 randomly generated watermarks. The watermark strength α was set to 0.15, and the

Image operation	Correlation
Blurring (in PaintShop Pro 4.12)	0.75
16% uniform noise (as in PSP 4.12)	0.95
Downsampling by a factor of 2	0.92
StirMark applied once	0.80
Unzign12 applied once	0.82

Table 1 Robustness with respect to image modifications.

watermark sequence was chosen for simplicity as $w_k = (-1)^k$. The correlation for 100 random watermarks is shown in Figure 5. The robustness with respect to JPEG compression was tested for quality factors from 5% to 85% and is shown in Figure 6. The robustness with respect to other image processing operations is summarized in Table 1. Both StirMark and Unzign were used with their default settings. Since Unzign12 cut the horizontal dimension of the image by 3 pixel values, the corresponding portion of the original image was used to bring the dimensions back to those of the watermarked image. Repetitive applications of StirMark did remove the watermark beyond detection. However, if the original image is available, the watermark can be easily detected even after multiple applications of StirMark. We used a simple motion vector estimator and resampled the image to correct for the geometrical deformation introduced by StirMark. The results are reported in Table 2.

Distortion	Correlation without adjustment	Correlation with adjustment
StirMark 2×	0.21	0.81
StirMark 3×	0.10	0.78
StirMark 3× + 20% uniform noise (as in PSP 4.12)	< 0.1	0.53
StirMark 3× + blurring 1× (as in PSP 4.12)	< 0.1	0.69
StirMark 3× + JPEG 25% quality compression	< 0.1	0.79
StirMark 3× + JPEG 15% quality compression	< 0.1	0.78

Table 2 Robustness after an adjustment for StirMark geometrical deformation.

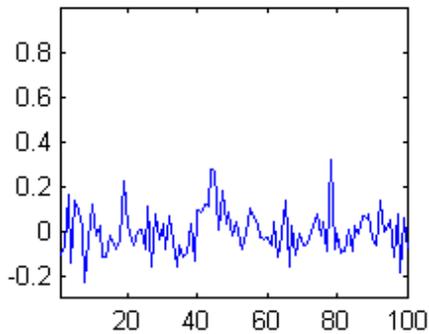


Figure 5 Correlation for 100 random watermarks.

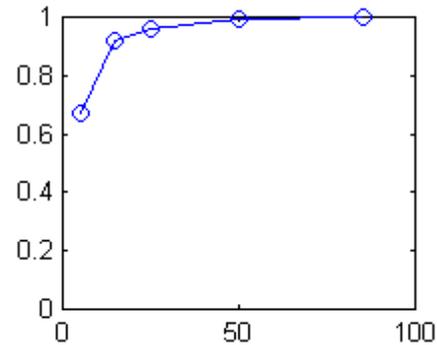


Figure 6 Robustness to JPEG compression.

To test the robustness with respect to the collusion attack, total of six images watermarked by different marks were averaged. The correlation coefficients were in the range from 0.51 to 0.71. The robustness experiments together with the test for correlations (Figure 5) between random watermarks suggest that a threshold of 0.4 should be used with this scheme. On the assumption that the projections are Gaussian distributed, this threshold gives the probability of false detection of the order of 10^{-4} .



Figure 7 Original image.



Figure 8 Watermarked image.

Test 2: Local scheme

In the local scheme, the image is divided into square subregions and each subregion is watermarked with a different set of orthogonal patterns. In our simulations, we used a 256×256 image of Lenna divided into 16 subregions of 64×64 pixels. The watermark strength was set to $\alpha=0.05$, and the watermark sequence was again $w_k=(-1)^k$. The watermark length was fixed at $J = 30$ to cut down on computing time. First, the original image was watermarked and then tested for presence of 100 randomly generated watermarks (Figure 9).

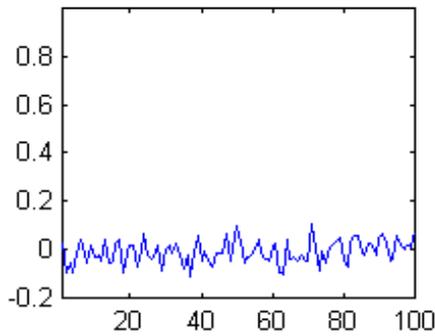


Figure 9 Correlation for 100 random watermarks.

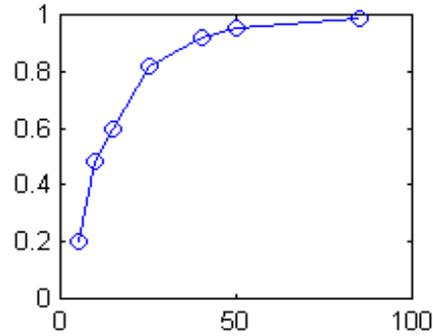


Figure 10 Robustness to JPEG compression.

Image operation	Correlation
Blurring (in PaintShop Pro 4.12)	0.68
16% uniform noise (as in PSP 4.12)	0.76
Downsampling by a factor of 2	0.53
2× downsampling, 16% uniform noise, 25% quality JPEG	0.47

Table 3 Robustness with respect to image processing operations.

The robustness with respect to JPEG compression is shown in Figure 10. By comparing the correlation values to random correlations in Figure 9, it appears that the threshold of 0.25 is appropriate in this case. Using this threshold, no false detections are produced for 100 random watermarks. The threshold enables a reliable detection of 10% quality (0.38bbp) compressed JPEGs. Other image processing operations, such as blurring, noise adding, and downsampling, and their combinations have been studied. A sample of the results is shown in Table 3. Further tests of robustness with respect to consecutive printing, copying, and scanning are currently undergoing.

4. Secure public black-box watermark detector

One of the most important arguments for using key-dependent basis functions is the fact that this concept may enable us to construct a secure public detector of watermarks that is implemented as a black-box in a tamper-proof hardware. Such watermark detectors will

find important applications in copy control of DVD [11–15]. The box accepts integer matrices on its input and outputs one bit of information. It is assumed that the complete design of the detector and the corresponding watermarking scheme are known except a secret key, and that an attacker has one watermarked image at his disposal. The latest attacks on public watermark detectors [11–15] indicate that it is not clear if a secure public watermark detector can be built at all. It has been proven that all watermark detectors that are thresholded linear correlators can be attacked using a variety of techniques [11–14]. Kalker [13,14] describes a simple statistical technique using which the secret key can be recovered in $O(N)$ operations, where N is the number of pixels in the image. The main culprit seems to be the fact that the quantities c_i that are correlated with the watermark sequence w_i can be directly modified through the pixel values, and the fact that the correlation function is linear. Linnartz and Cox [11,12] attack public detectors by investigating the sensitivity of the watermark detector to individual pixels for a critical image – the image at the detection threshold. Once the most influential set of pixels is found, its gray levels are scaled and subtracted from the watermarked image. They repeat the process in a hope to converge to an image that does not have the watermark. The assumption here is that we can actually learn the sensitivity of the detection function *at the watermarked image* from its sensitivity *at the critical image* that will generally be far from the watermarked image.

In order to design a watermarking method with a detector that would not be vulnerable to those attacks, we need to mask the quantities that are being correlated so that we cannot purposely change them through pixel values and we must introduce nonlinearity into the scheme to prevent the attack by Linnartz and Cox [11,12]. Towards this purpose, we propose to use key-dependent basis functions and a special nonlinear index function. This technique will be described in more detail in a forthcoming paper. The watermarking technique will work on the same principle as before: the watermark sequence $w_i \in \{-1, 1\}$ is embedded into an image by adjusting the projections $c_i, i=1, \dots, J$ of the image onto the orthogonal patterns so that $ind(c_i)=w_i$ for a carefully chosen index function $ind(x)$. The index function is a continuous function similar to $\sin(x)$ with an increasing wavelength. It plays the role of a quantization-like function. We propose the following function

$$ind(x) = \sin\left(\frac{\pi}{\ln(q)} \ln\left(\frac{x}{x_0}\right)\right),$$

where $q = (1+\alpha)/(1-\alpha)$, $x_0=1$. This function has the following properties: (i) any $x \geq 1$ can be modified by at most $2\alpha\%$ in order to change its index $ind(x)$ from any value to either 1 or -1 . By embedding a watermark into the projections, most of them will be modified by a small value, but some can be modified by almost $2\alpha\%$.

To detect the watermark sequence w_i in image I , the watermark detector first projects the image I onto the secret patterns f^i , calculates the values of the correlation, applies the index function and correlates the result with the watermark sequence w_i :

$$D(I) = H\left(\sum_{i=1}^J w_i ind(c_i) - Th\right),$$

where Th is the detection threshold, $H(x)$ is the step Heaviside function, and $D(I)$ is the detection function applied to I .

Now we need to argue why this scheme may not be vulnerable to previously described attacks. First of all, since c_i are projections on unknown patterns, one cannot purposely change the pixels values – the input of the detector. The relationship between the projections and pixel values is unknown. If we were able to calculate c_i from the pixel values, we could learn the watermark values w_i from cleverly chosen perturbations. Second, the sensitivity of the detector function at a critical image C (or, equivalently, the values of partial derivatives with respect to pixel values) cannot be directly related to sensitivity values at the watermarked image. By changing the pixel g_{rs} by Δ , we can express the corresponding change in the detector function as

$$\Delta D(C) = \Delta \sum_{i=1}^J w_i f^i_{rs} ind'(c_i),$$

where f^i_{rs} is the gray level of the (r, s) -th element of the i -th pattern. What we can learn from sensitivity analysis at the critical image is the value of the summation. However, this value depends on the unknown parameters f^i_{rs} and on the values of the derivative of the index function at the projections c_i corresponding to the critical image. However, the projections of the critical image and the original watermarked image will generally be very different. This indicates that it may be rather difficult to utilize the leakage of information gained by perturbing the critical image.

Preliminary tests of the robustness of this scheme suggest that it has extremely good robustness with respect to filtering, JPEG compression, and resampling. More detailed theoretical investigation and experiments are needed, however, before this scheme can be termed as a successful solution to the public watermark detector. Detailed analysis of the proposed scheme and a watermark detector will be the subject of further research.

5. Conclusions and future directions

In this paper, we introduced the concept of key-dependent basis functions and described how it can be used for designing secure robust watermarking schemes and secure public black-box watermark detectors. The new schemes overcome a possible security weakness of global, non-adaptive schemes that apply watermark patterns spanned by a small number of publicly known basis functions. The watermark is embedded into the projections of an image onto the secret set of key-dependent patterns. The robustness of the watermarking scheme with respect to filtering, lossy compression, and combinations of other attacks was studied. Finally, we proposed a candidate for a watermarking scheme that has a secure public watermark detector.

Future research will include further study of the robustness of the new scheme with respect to image distortions. One important future research direction is the development of secure public black-box watermark detectors using key-dependent basis functions. It appears that the concept of key-dependent basis functions together with special quantization index functions leads to very robust watermarking schemes for which the

construction of a secure public black-box watermark detector is possible. Most importantly, we plan to rigorously estimate the complexity of possible attacks on the public detector.

Acknowledgments

The work on this paper was supported by Air Force Research Laboratory, Air Force Material Command, USAF, under a Phase I SBIR grant number F30602-97-C-0209. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Air Force Research Laboratory, or the U. S. Government.

References

- [1] I.J. Cox, J. Kilian, T. Leighton, and T. Shamoan, "Secure Spread Spectrum Watermarking for Multimedia," NEC Research Institute, *Technical Report* 95-10.
- [2] I.J. Cox and M.L. Miller, "A review of watermarking and the importance of perceptual modeling", *Proceedings of Electronic Imaging '97*, February 1997.
- [3] M.D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia Data Embedding and Watermarking Technologies", *Invited Paper, to appear in the Proceedings of the IEEE*, 1998.
- [4] A.H. Tewfik, M.D. Swanson, B. Zhu, K. Hamdy, and L. Boney, "Transparent Robust Watermarking for Images and Audio." *IEEE Trans. on Signal Proc.*, 1996.
- [5] J.-F. Delaigle, C. De Vleeschouwer, B. Macq, "Digital watermarking of images," *Proceedings of the IS&T/SPIE Symposium on Electronic Imaging Science and Technology*, 1996.
- [6] N. Jayant, J. Johnston, and R. Safranek, "Signal Compression Based on Models of Human Perception", *Proceedings of the IEEE*, Vol. 81, No. 10, Oct 1993.
- [7] N. Jayant, J. Johnston, and R. Safranek, "Perceptual Coding of Images", *SPIE Vol. 1913*, 1993.
- [8] B. Zhu and A.H. Tewfik, "Low Bit Rate Near-Transparent Image Coding", *SPIE Vol. 2491*, 1995.
- [9] B.O. Comiskey and J.R. Smith, "Modulation and Information Hiding in Images," in: *Information Hiding, First International Workshop*, edited by Ross J. Anderson. Cambridge, U.K., May 30–June 1, 1996, Proceedings. Lecture Notes in Computer Science, Vol. 1174, Springer-Verlag, 1996.
- [10] Hartung and B. Girod, "Digital Watermarking of Raw and Compressed Video", *Proc. European EOS/SPIE Symposium on Advanced Imaging and Network Technologies*, Berlin, Germany, Oct. 1996.
- [11] I.J. Cox and Jean-Paul M.G. Linnartz, "Public watermarks and resistance to tampering", in *Proceedings of the ICIP*, October 1997, CD version of Proceedings.
- [12] I.J. Cox and Jean-Paul M.G. Linnartz, "Some general methods for tampering with watermarks", *preprint*, 1998.

- [13] T. Kalker, “Watermark Estimation Through Detector Observation”, Philips Research Eindhoven, Netherland, *preprint* 1998.
- [14] T. Kalker, J.P. Linnartz and M. van Dijk, “Watermark Estimation Through Detector Analysis”, preprint submitted to *ICIP-98*.
- [15] J.P. Linnartz and M. van Dijk, “Analysis of the sensitivity attack against electronic watermarks in images”, in *Proceedings of the Workshop on Information Hiding*, Portland, April 1998, submitted.
- [16] S. Craver, N. Memon, B.-L. Yeo, and M. 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.
- [17] M.G. Kuhn, *StirMark*. Available at <http://www.cl.cam.ac.uk/~mgk25/stirMark/>, Security Group, Computer Lab, Cambridge University, UK (E-mail: mkuhn@acm.org), 1997.
- [18] *Unzign*. Available at <http://altern.org/watermark/> (E-mail: unzign@hotmail.com), 1997.