HYBRID ROBUST WATERMARKING RESISTANT AGAINST COPY ATTACK

Frédéric Deguillaume, Sviatoslav Voloshynovskiy, and Thierry Pun
CUI, University of Geneva,
24 rue du Général Dufour, CH 1211 Geneva 4, Switzerland
e-mail: {Frederic.Deguillaume,svolos,Thierry.Pun}@cui.unige.ch

ABSTRACT

Three major aspects in the protection of digital documents have been pointed out, each of them with different requirements: first copyright protection, protecting ownership and usage rights; secondly tamper-proofing, aiming at checking document integrity; and thirdly authentication, the purpose of which is to check the authenticity of a document. While robust watermarks are typically used for copyright protection, fragile or semi-fragile watermarks are proposed to solve tamper-proofing and authentication. However both concepts are generally proposed separately; further, most of robust watermarking schemes are vulnerable to the copy attack, which allows copying a watermark from one document to another without need for any apriori knowledge [4]. In this paper we propose an hybrid watermarking method which joins a highly robust watermark with a fragile/semi-fragile watermark, combining copyright protection and tamper-proofing, and being at the same time resistant against copy attack; the fragile information is inserted in a way which fully preserves the robust part.

1 INTRODUCTION

Numerous security threats concerning the rapidly growing multimedia market based on digital technologies have been pointed out, the first to be identified being the ease with wich unauthorized copies could be made. Therefore, robust watermarking for copyright protection has been proposed, with the two main requirements of high *robustness* (the watermark should survive any kind of malicious or unintentional alterations), and low *visibility* (the watermark should not introduce perceptible artifacts). Another requirement is that the scheme should be *oblivious* (the original image not needed for the extraction process).

Robust watermarking schemes have been proposed, consisting in either spatial domain, or transform domain watermarks. One of the main issue was the resistance to geometrical distortions which desynchronize the watermark and make it unreadable. Solutions against geometrical transform use either a transform invariant domain watermak like the Fourier-mellin transform, or an additional template for resynchronization, or a self-reference watermark based on the Autocorrelation Function (ACF) of a repetitive watermark [3]. Self-reference watermarks have been shown to have as main advantage over other methods the fact that they exploit the redundancy of the regular structure of the watermark in order to robustly estimate the undergone geometrical distortions. We previously proposed a method based on this concept known as *Berkut 1.0*, which is robust to general affine transforms [7] as well as to non-linear distortions and to the random bending attack (RBA) [8]; our approach uses the ACF or magnitude spectrum of a periodical watermark, at the global level to recover from affine transforms, and at the local level to recover from RBA.

Another important threat with respect to multimedia document is the ease offered by today technologies for tampering or counterfeiting. Digital cameras are constantly growing in quality while becomming widely available, and softwares such as Paintshop Pro make it very easy to perform complex modifications without visible artifact. Although this is usefull for artistic applications, it is a serious problem for legal applications such as evidences in trials, for insurances, in medical imaging, etc. Classical analysis techniques used for authenticating analog photographs are ineffective. Of course a global cryptographic signature detects tampering, but is unable either to highlight which areas have been modified, or to assess the severity of the alteration. Therefore the solution is again watermarking, used here to attach check-codes of local areas inside the image itself. As such watermarks do not need the same level of robustness than for copyright protection, they can be either *fragile* (any modification, even slight pixel change, are detected), or *semi-fragile* (which offers a tolerance to some "acceptable" alteration such as low-level compression or slight contrast adjustment).

Generally, the image is first divided into small blocks for locality, a key-dependent hash function applied to each of them, and the hash-codes embedded to their corresponding blocks, usually in the least significant bits

(LSB) of pixels. Tampering is then detected where recomputed codes do not match the stored codes. Wong [6] proposed such a blockwise approach. Semi-fragile watermarks are more tolerant, and can even measure the severity of the alteration; a robust watermarking scheme can be used for this, however this approach is insecure since robust watermarks are usually additive, making them vulnerable to the copy attack: the signal can be easily estimated using denoising techniques and copied to another image [4]. Another possibility is to compute *robust* hashes which are tolerant to slight modifications, and to embed them robustly.

Watermarking methods above are either robust schemes, or fragile/semi-fragile schemes; however approaches combining both for copyright and tamperproofing/authentication application are rarely proposed. Fridrich [1] proposed such an hybrid method, but uses a watermark with relatively low robustness. Therefore we propose: first, to join a highly robust watermark with a fragile watermark for combined copyright protection, tamperproofing and authentication (section 2); secondly, a smart embedding of the fragile part which fully preserves the robust watermark, as described in the same section; thirdly, an extension of this scheme to a joint-robust and semi-fragile watermark (section 4). Section 3 outlines the security aspects of block-wise hash-coding. Section 5 shows the ability for tamper detection of our poposal, and demonstrates that our approach is resistant to the copy attack.

2 HYBRID WATERMARKING

We propose, in the prototype *Berkut 2.0*, to join our previously developped higly robust self-reference watermarking technology [7, 8] with a blockwise fragile scheme based on cryptographically secure hash-codes similar to Wong's approach [6]. The robust component w is achieved by tiling a block containing the information over the whole cover image x, resulting in a periodical watermark which is added to produce the stego image y. At the extraction step, the ACF or the magnitude spectrum of the estimated watermark $\hat{w}$ (from the possibly attacked stego image y') results in a regular grid of local maxima, or peaks, allowing a robust estimation and compensation of the undergone affine distortion. Non-linear distortions (such as RBA) are considered at the local level. Visual masking is ensured by modulating the watermark by a Noise Visibility Function (NVF), a modulation transfer function (MTF) of the human visual system (HVS) in function of the resolution of the image based on a wavelet transform (WT). Robustness to any signal processing attack is obtained by further encoding the copyright message based on error correction codes (ECC) whith soft decoding, such as $low-density\ parity-check\ codes$ (LDPC) codes or *turbo codes*; these codes are known to closely approach the theoretical maximum performance, and use simple iterative decoding algorithms.

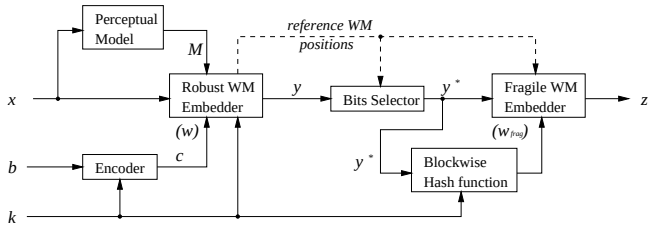

Figure 1: Hybrid embedding: first the robust watermark is embedded, then the fragile watermark.

2.1 Hybrid watermark embedding

One robust watermark block further consists in two non-overlapping, i.e. orthogonal, components: an *informative* watermark w_{inf} holding the copyright message b, encrypted with a secret key k; and a *reference* watermark w_{ref} only depending on k, used as a pilot for translation/cropping determination and for other side information usefull for the decoding step.The allocated positions in block also depend on k. Since the watermark further has a density less than 1, positions in which no information is embedded still remain.

The fragile watermark w_{frag} is based on key-dependant (from the secret key k) blockwise cryptographically secure hash-codes. w_{frag} blocks may or may not coincide with w blocks (actually fragile block may be subblocks from robust blocks for better locality in the tamper detection). An important issue is to preserve the original robustness of the robust scheme: first, embedding the fragile part into the LSB of selected pixels ensures very low modification; secondly, we embed the fragile watermark in selected positions belonging to the reference watermark only (and possibly also to position containing no watermark). Thus w_{inf} is untouched, and on average at most 50% of positions in w_{ref} are altered by $+1$ or -1. Since w_{ref} usually covers not more than 20% of the area of w, this makes w_{frag} and w almost orthogonal. At the same time the visual impact of the fragile part is much lower than the visual impact of the robust part. The block diagram 1 shows the hybrid embedding process. Obviously, the fragile component has to be applied *after* the robust one, and hash-codes should take as input $y^\star$, a version of y where all bits selected to embed w_{frag} are cleared to 0 by the "bits selector" block. The stego image containing both robust and fragile watermark is z.

2.2 Hybrid watermark extraction

The block diagram 2 shows the extraction and authentication part. The robust extractor estimates the robust watermark $\hat{w}$ from the possibly attacked and tampered stego image z', and decode an estimation of the copyright message $\hat{b}$; if necessary the applied geometrical distortions are compensated inside this part. The authentication part takes z' as input; computes estimated

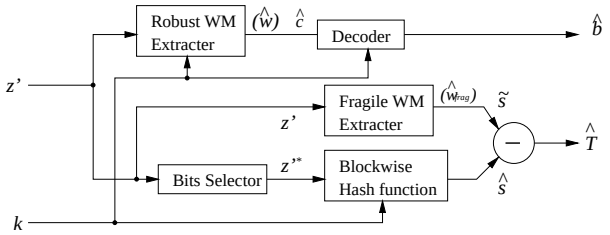

Figure 2: Hybrid extraction: the robust and the fragile watermark are extracted separately, afterwards a decision have to be taken.

signatures $\hat{s}$ from $z'^{\star}$ (a version of z' where the LSBs used for w_{frag} have been set to 0); extract w_{frag} from z' and get the embedded signatures $\tilde{s}$; and finally output a tamper map $\hat{T}$ by comparing $\tilde{s}$ and $\hat{s}$. At the end the generic following decision can then be made, depending of the targeted application:

- $\hat{b}$ **is correctly decoded and** $\hat{T}$ **is fully matched:** we fully authenticate the image.
- $\hat{b}$ **is correctly decoded but** $\hat{T}$ **failed:** if $\hat{T}$ shows local tampering, malicious modification probably occured: we authenticate the image partially and point out modified regions; if $\hat{T}$ shows global tampering, we fully reject the image, copyright violation probably occured.
- **both** $\hat{b}$ **and** $\hat{T}$ **are failed:** fully reject the image as non authentic, but no copyright was pointed out.

Note that a copy attack correspond to the second item, since it would make $\hat{b}$ still correctly decoded but $\hat{T}$ globally failed: therefore by rejecting this case, our hybrid approach is resistant to copy attack.

3 SECURITY OF FRAGILE WATERMARK

It has been noticed that schemes based on the hashing of independant blocks like Wong's scheme were vulnerage to various tampering attacks such as substitutions attacks [5]. Such attacks are mainly made possible first due to the independance of blocks, and secondly as a result of the *anniversary paradox* affecting hash functions used as signatures. An attack based on this paradox uses the fact that for hash-codes of n bits, the probability that two different blocks result in the same code is already equal to 50% when only $\sqrt{n}$ different blocks are gathered: therefore for hash-codes of 64 bits, only 2^{32} block samples are needed to obtain a collision; the availability of large databases of images protected with the same key would make this attack realistic. Among solutions proposed to fix these security problems, the main ones consists in [5]: using hash-codes of 128 bits or more to defeat the anniversary attack; making hash-codes embedded into one block also dependent from neighbouring blocks, or using overlapping blocks as input of hash-codes and smaller non-overlapping blocks for embedding these codes; using non-deterministic cryptographic signatures to defeat more sophisticated attacks, and making hash-codes also dependent from the hash-codes of neighbouring blocks. Finally we propose to include the 8 neighbouring blocks as well as the coordinates of the current block in the input of hashes, and to use non deterministic signatures.

4 SEMI-FRAGILE EXTENSION

Note that for the second item, when $\hat{b}$ is correctly estimated but $\hat{T}$ completely failed, we have not way to distinguish an innocent alteration such as compression from bulk tampering. Here we propose a generic approach to extend our method to a more flexible one: first we use a semi-fragile watermark, achieved by a robust image hashing. Unlike previous hash functions, the hash value of robust hashing changes continously with the input, by changing only a small number of bits for a perceptually slight modification of the input. Robust hashing, for example based on a low-pass transform [2], can be used. Secondly, we have to replace the LSB approach by a robust embedding. The hash-codes can then be embedded as a payload within the robust watermark, maybe with a weaker strength, and/or within the reference watermark; positions independent from the robust payload can still be chosen.

However this semi-fragile extension, unlike the original approach, has two main drawbacks: first, finding visual hashes which are at the same time robust enough and secure is still an open problem; secondly, embedding hash-codes as a payload requires to reduce the density of the robust watermark, and to increase the size of blocks: the robustness of the information would probably be decreased (unlike with the original approach), and less locality would be achieved in the detection of tampering. We propose therefore two variants, depending on the targeted application: first, a *strict* hybrid watermarking as described in the previous sections - based on strict hash functions and fragile LSB embedding; secondly, a *tolerant* hybrid watermarking, using robust or visual hash-codes and robust embedding within the joint robust watermarking. While the strict variant allows to authenticate digital documents only, the tolerant variant would be mostly suitable in the case of digital/analogic conversion - it can be used for hard-copy documents such as bank-notes or value papers.

5 RESULTS

Two experiments have been performed, using the strict version of our method. First, we embedded the hybrid watermark into the image "leysin", and then tried to detect it: the robust watermark was sucessfully decoded, and the image fully authenticated. Then "leysin" was maliciously tampered: the robust watermark was still decoded, but modified regions where detected; figure 3 shows this attack. Finally we tested the method with

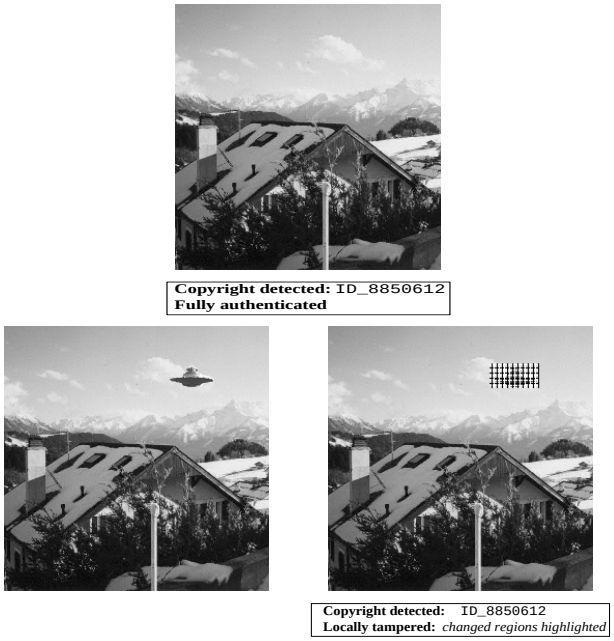

Figure 3: Local tampering experiment. Top: watermarked "leysin". Bottom left: tampered watermarked "leysin". Bottom right: local tampering have been detected.

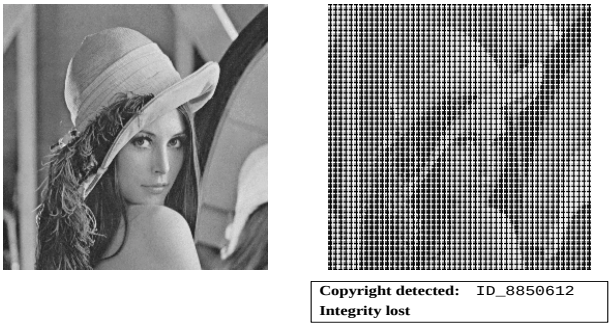

Figure 4: Copy attack experiment. Left: watermartk was copied from "lena" to "leysin". Right: robust watermark is decoded, but image integrity has been completely lost.

copy attack. The watermark from "leysin" has been copied to "lena"; while the copyright message was still decoded, the image was completely rejected as shown by figure 4. Simultaneously, we have tested our method according to the Stirmark benchmark (table 1).

6 CONCLUSIONS

In this paper we presented an hybrid watermarking scheme combining copyright protection and detection of tampering. For this purpose we used the highly robust watermarking scheme we previously developped, and a fragile watermark based on local hash-codes. The fragile information does not decrease the robustness of the robust part, and the scheme is resistant to the copy attack. The approach is suitable for joint copyright protection/tracking and tamperporoofing/authentication purpose.

Table 1: Averaged results of system perfomance according to Stirmark3.1.

Stirmark attack	*Averaged score*
Signal enhancement	1,00
Compression (JPEG/GIF)	0,99
Scaling	1,00
Cropping	0,99
Shearing	1,00
Rotation (auto-crop, auto-scale)	0,99
Column and line removal	1,00
Flip	1,00
Random Bending Attack	1,00

References

[1] J. Fridrich. A hybrid watermark for tamper detection in digital images. In *ISSPA'99 Conference*, Brisbane, Australia, August 1999.

[2] J. Fridrich. Visual hash for oblivious watermarking. In *IS&T/SPIE Proceedings*, volume 3971, San Jose, California, USA, January 2000.

[3] M. Kutter. *Digital image watermarking: hiding information in images*. PhD thesis, EPFL, Lausanne, Switzerland, August 1999.

[4] M. Kutter, S. Voloshynovskiy, and A. Herrigel. Watermark copy attack. In Ping Wah Wong and Edward J. Delp, editors, *IS&T/SPIE's Electronic Imaging 2000*, volume 3971 of *SPIE Proceedings*, San Jose, California USA, 23–28 jan 2000.

[5] Hae Yong Kim Paulo S. L. M. Barreto and Vincent Rijmen. Toward a secure public-key blockwise fragile authentication watermarking. In *In IEEE ICIP2001*, pages pp. 494–497, Thessaloniki, Greece, October 2001.

[6] P.W.Wong. A public key watermark for image verification and authentication. In *in Proceedings of IEEE Int. Conf. on Image Processing, vol. 1, MA11.07*, 1998.

[7] F. Deguillaume S. Voloshynovskiy and T. Pun. Content adaptive watermarking based on a stochastic multiresolution image modeling. In *European Signal Processing Conf. EUSIPCO'2000*, Tampere, Finland, September 2000.

[8] Frédéric Deguillaume Svyatoslav Voloshynovskiy and Thierry Pun. Multibit digital watermarking robust against local nonlinear geometrical distortions. In *In IEEE ICIP2001*, pages pp. 999–1002, Thessaloniki, Greece, October 2001.