TOWARD GENERIC IMAGE DEWATERMARKING?

C. Rey, G. Doërr, G. Csurka, and J.-L. Dugelay

Dept. of Multimedia Communications
Institut Eurécom, Sophia Antipolis, France
http://www.eurecom.fr/~image

ABSTRACT

A significant effort has been put in designing watermarking algorithms during the last decade. But today, the watermarking community needs some advanced attacks and fair benchmarks in order to compare the performances of different watermarking technologies. Moreover attacks permit to find the weaknesses of an algorithm and consequently trigger further research in order to overcome the problem. This state of mind motivates the creation of the European Certimark project.

After a short definition of the keyword dewatermarking, we present an original attack based on self similarities. This attack is then put to the test with three different publicly available watermarking tools. Finally we shortly discuss the feasibility of a generic attack i.e. a dewatermarking attack which should succeed in removing whatever watermark inserted by whatever watermarking tools.

1. INTRODUCTION

Image watermarking is now a major domain. Basically, digital watermarking allows owners or providers to hide an invisible and robust message inside multimedia content, often for security purposes, in particular owner or content authentication. There exists a complex trade-off between three parameters in digital watermarking: capacity, visibility and robustness. Robustness means that the retriever is still able to recover the hidden message even if the watermarked content has been altered after embedding. Today, most of the proposed watermarking schemes are robust against normal processing e.g. low pass filtering, JPEG compression. However most of them are still weak against malicious attacks.

From the beginning, a sort of competition between *attackers* and *watermarkers* has existed. Nevertheless, research due to attackers benefits the entire watermarking community. As soon as a new attack is designed, watermarkers try to improve their algorithms in order to survive this new attack, often via a preventive procedure. Moreover it is necessary to develop attacks in order to set up benchmarks which will allow a fair comparison between the different proposed watermarking schemes. Stirmark [7] is currently considered as one of the most efficient malicious attack. It is mainly based on random local geometric distortions (hard to prevent or to compensate) of the cover that most often traps the synchronization between the encoder and the decoder. But the watermark is still present and there is no guarantee for the attacker that a possible future improved version of the decoder will not solve the problem.

In the present paper, we present an original attack which is assumed to definitely remove the watermark. In Section 2, we specify the basic requirements that an attack should meet in order to be considered as a dewatermarking attack. In Section 3, we present our approach for still images based on self similarities. In Section 4, we show the performances of our attack against three publicly available watermarking tools. Finally we bring the feasibility of a generic dewatermarking attack up for discussion in Section 5.

2. IMAGE DEWATERMARKING

The keyword *dewatermarking* is partially self-explanatory by analogy with denoising, even if it is not yet commonly used in the literature. It means that the attack should not leave any underlying evidence of the presence of the watermark. It is radically different from a desynchronization attack like Stirmark. When an attacker hacks a large database, he does not want to get caught later because a new version of the detector is not trapped any more by his attack. He wants to be sure that any copyright information has been removed once for all.

Obviously, the ideal dewatermarking attack would consist of blindly restoring the original document from the watermarked one. But such a perfect attack is quite impossible to implement in practice. As a result, by dewatermarking, we mean an attack that fulfills the following specifications:

1. The detector is no longer able to recover the watermark.

2. The computation of a quantitative measure of distor-

Contact author: jld@eurecom.fr. This work has been supported by the Certimark [2] project.

tion, e.g. PSNR or wPSNR [9], between the watermarked document and the document resulting from the malicious manipulation remains pertinent i.e. the attack introduces no geometric distortion in order to remain compliant with the recent modeling of the attack channel [6].

3. The attack should introduce a fair additional distortion. The distance between the watermarked and the attacked documents should be close (or even inferior) to the distance existing between the original and the watermarked documents. That is to say, the distance between the watermarked and the attacked documents is less than twice the distance between the original and the watermarked documents.

4. The attack should insure that a future improved version of the decoder alone cannot overcome the problem. The protection of the documents are definitely lost and technology providers have to rework both embedder and retriever.

Obviously, many traditional image processings (filtering, lossy compression) can be classified as dewatermarking attacks if they succeed to remove the watermark and some recent attacks [8] already fulfill those requirements.

3. APPROACH FOR STILL IMAGES

Our dewatermarking attack for still images basically exploits self-similarities of the image. Self similarities can be seen as a particular kind of redundancy. Usually correlation between neighbor pixels is taken into account. With self similarities, it is the correlation between different parts (more or less spaced) of the image which is of interest. This idea has already been used with success for fractal compression [3].

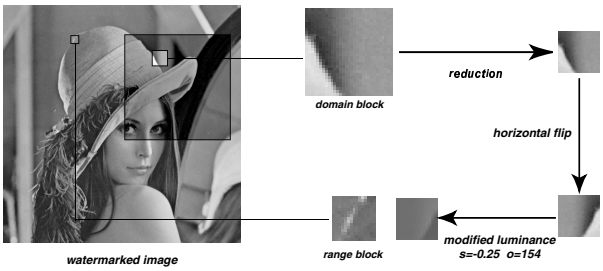

Fig. 1. Self similarities process

The basic idea of the attack consists in substituting some parts of the picture with some other parts of itself (or even from an external codebook) which are, or look, similar. This process is depicted in Figure 1 and explained in the next subsection. The objective is to approximate, to stir the watermarked signal while keeping clear the cover signal. Even if self similarities can be realized in various transform domain (DCT [1], wavelet), we restrict our presentation here to the attack in the spatial domain.

3.1. Attack in the spatial domain

In the spatial domain, the original image is scanned block by block. Those blocks are labeled *range blocks* (block $\mathbf{R_i}$) and have a given dimension $n \times n$. Each block $\mathbf{R_i}$ is then associated with another block $\mathbf{D_i}$ which looks similar (modulo a pool of possible photometric and geometric transformations) according to a Root Mean Square (RMS) metric defined by the following formula:

$$RMS(f,g) = \frac{1}{n}\sqrt{\sum_{x=1}^{n}\sum_{y=1}^{n}\Big(f(x,y) - g(x,y)\Big)^2} \qquad (1)$$

The block $\mathbf{D_i}$ is labeled *domain block* and is searched in a codebook containing Q blocks $\mathbf{Q_j}$. Those blocks may be blocks from the same image or from an external unwatermarked database. In practice, for a given range block $\mathbf{R_i}$, a window is randomly selected in the image. The blocks belonging to this window provide the codebook. Each block $\mathbf{Q_j}$ is scaled if needed in order to match the dimensions of the range block $\mathbf{R_i}$. A set of $T_\mathbf{k}$ geometrically transformed blocks $T_\mathbf{k}(\mathbf{Q_j})$ is then built (identity, 4 flips, 3 rotations). For each transformed block $T_\mathbf{k}(\mathbf{Q_j})$, the photometric scaling $\mathbf{s}$ and offset $\mathbf{o}$ is computed by minimizing the error between the transformed block $g = T_\mathbf{k}(\mathbf{Q_i})$ and the range block $f = \mathbf{R_i}$ by the Least Mean Square method.

$$R = \sum_{x=1}^{n}\sum_{y=1}^{n}\Big(\mathbf{s}.g(x,y) + \mathbf{o} - f(x,y)\Big)^2 \qquad (2)$$

Eventually, the transformed block $\mathbf{s}.T_\mathbf{k}(\mathbf{Q_j}) + \mathbf{o}$ which has the lowest RMS distance with the range block $\mathbf{R_i}$ is found and the corresponding block $\mathbf{Q_j}$ will be the domain block $\mathbf{D_i}$ associated with the range block $\mathbf{R_i}$. Since the two blocks $\mathbf{R_i}$ and $\mathbf{D_i}$ looks similar, we can substitute $\mathbf{R_i}$ with the transformed version of $\mathbf{D_i}$. As a result, the image will be slightly modified but the watermark signal will be randomly spread through the image and the detector will be unable to retrieve it.

3.2. Additional specifications

Self similarities were not designed for dewatermarking. In this case a perfect reconstruction is not expected. In fact a minimum error during the block association is even needed so that the watermark is removed. As a result, a threshold τ has been introduced and the original rule to associate a

domain block with a range block has been modified. Now, for each range block, we search for the transformed block $\mathbf{s}.T_k(\mathbf{Q_j}) + \mathbf{o}$ which has the lowest RMS distance with the range block $\mathbf{R_i}$ *above the threshold* τ. If all the RMS distances are below the threshold, the block with the greatest distance is kept. In order to have an image dependent threshold, it is chosen in such a way that a given percentage p of the range blocks are not optimally substituted. As a result, two IFS iterations are needed. In the first iteration, the threshold is set to zero and the cumulative histogram of the errors between the range blocks and the domain blocks is built. The adaptive threshold is then determined in order to interfere with p percents of the substitutions during the second iteration.

This new specification is likely to introduce visible artifacts. In order to prevent this effect, two constraints have been added:

- Only a given part of the domain block is substituted with the range block. In our case, we used a circular mask inscribed in the block.
- Overlapping range blocks have been used. Consequently, specific care must be taken during the reconstruction. A simple substitution is not any more pertinent. Instead the domain blocks are accumulated in a temporary image and, at the end, each pixel value is divided by the number of blocks that contribute to the value of this pixel.

4. EXPERIMENTAL RESULTS

This attack has been tested with three publicly available watermarking tools that offer roughly the same capacity (a few bits). A wide range of color images have been tested, although we only report the results with *lena* in this article. Moreover, we made the assumption that the attacker knows in which color channel is embedded the watermark. Indeed, even if this hint is kept secret, it is fairly easy to guess.

Our attack has been tested against D******* in a first experiment. The watermark seems to be mainly embedded in the V channel of the HSV color space. We find out that around 60% of the block associations need to be disturbed in order to remove the watermark in all the tested images. This results in a quite good image quality as it can be seen in Figure 2. Visually one can notice that the textured areas are a little bit affected. The PSNR (resp. wPSNR)[1] is equal to 40.32 dB (resp. 53.90 dB) between the original image and its watermarked version, while it is equal to 35.67 dB (resp. 51.54 dB) between the watermarked image and its attacked version. As a result, this attack can be considered successful.

[1]The PSNR and the wPSNR are computed on the Y channel only of YUV color space.

(a) Original image

(b) D*******

(c) Attacked image

Fig. 2. Attack against D*******.

In a second experiment, S***I** has been put to the test. The watermark is strongly embedded in the B channel of the RGB color space [4]. In order to face the strength of the watermark, we need to disturb 99% of the block associations. It results in a strong degradation of the blue channel. But this degradation is quite invisible since the human eye is less sensible to the blue channel as it can be seen in Figure 3 which shows the luminance of the attacked image. The PSNR (resp. wPSNR) is equal to 49.05 dB (resp. 59.73 dB) between the original image and its watermarked version, while it is equal to 46.52 dB (resp. 59.24 dB) between the watermarked image and its attacked version. Once again,

(a) S***I**

(b) Attacked image

Fig. 3. Attack against S***I**.

the attack is a success.

(a) S***S*** (b) Attacked image

Fig. 4. Attack against S***S***.

In the last experiment, we tested S***S***. The watermark seems to be mainly embedded in the channel Y of the color space YUV. It has been determined experimentally that 92% of the block associations have to be disturbed in order to remove the watermark. This results in strong visible artifacts as can be seen in Figure 4. At least for the moment, the attack is a failure.

5. CONCLUDING REMARKS

We have described in this paper an efficient dewatermarking attack. This attack has been partially integrated into Stirmark benchmark v4.0 [7] and we expect that it will provide a useful tool for testing watermark robustness. It fulfills the requirements specified earlier and succeeds in trapping two out of the three investigated watermarking schemes. However, even if all the proposed attacks have a common root (self similarities), the parameters of the attack differ (attacked color channel, percentage p). It seems indeed difficult to build a generic dewatermarking attack. This is due to the high specialization of the watermarking technologies. Defeating one watermarking algorithm does not mean the others will be defeated. For example, a simple averaging filter of width 5 usually removes the watermark inserted by D*******. On the other hand, it will leave the watermarked inserted by S***I** or S***S*** unaffected! Anyway, having a pool of dedicated attacks is not completely useless.

Recently some researchers found some exciting results in steganalysis [5]. The authors showed that it is possible to predict if an image has been watermarked and by which technology. So now we have a toolbox containing multiple simple attacks optimized for a single technology in one hand, and an oracle which is able to say which watermarking technology has been used in the other hand. Combine those two items together and you obtain a very powerful tool for attackers. We can now make a straightforward analogy with an anti-virus software. For any new incoming watermarking technology (the virus), the attackers only have to design a simple dewatermarking attack (the anti-virus) and to update the oracle. As a result, if an attacker does not want to get caught, he just has to keep his system up to date.

This attack will be further investigated in the future. On one hand, a possible extension of this attack in the wavelet domain will be studied. On the other hand, steganalysis studies will be launched in order to try to automate the detection of the color channel to attack.

6. REFERENCES

[1] K.-U. Barthel, J. Schüttemeyer and P. Noll "A New Image Coding Technique Unifying Fractal and Transform Coding", in *IEE on Image Processing*, Austin, USA, November 13-16 1994.

[2] Certimark, http://vision.unige.ch/certimark

[3] Y. Fisher, *Fractal Image Compression: Theory and Application*, editor Springer-Verlag, New York, 1995.

[4] M. Kutter, F. Jordan and F. Bossen, "Digital Signature of Color Images Using Amplitude Modulation", in *Proceedings of Electronic Imaging*, San Jose, USA, February 1997.

[5] N. Memon I. Avcibas and B. Sankur, "Steganalysis of Watermarking Techniques Using Image Quality metrics", in *Proceedings of SPIE Security and Watermarking of Multimedia Contents III*, San Jose, USA, January 22-25 2001, vol. 4314.

[6] P. Moulin and J. O'Sullivan, "Information Theoretic Analysis of Information Hiding", *Preprint*, September 1999.

[7] F. Petitcolas, R. Anderson and M. Kuhn, "Attacks on Copyright Marking Systems", in *Proceedings of Information Hiding*, Portland, USA, April 15-17 1998. http://www.cl.cam.ac.uk/~fapp2/watermarking

[8] K. Tsang and O. Au, "A Review on Attacks, Problems and Weaknesses of Digital Watermarking and the Pixel Reallocation Attack", in *Proceedings of SPIE Security and Watermarking of Multimedia Content III*, San Jose, USA, January 22-25 2001, vol. 4314.

[9] S. Voloshynovskiy, A. Herrigel, N. Baumgaertner and T. Pun, "A Stochastic Approach to Content Adaptive Digital Image Watermarking", in *Third International Workshop on Information Hiding*, Dresden, Germany, September 29 - October 1 1999.